ACCELERATED BREGMAN PROXIMAL GRADIENT METHODS FROM DUAL GEOMETRIC PERSPECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study Bregman proximal gradient (BPG) algorithms under relative smoothness for convex, relatively strongly convex, and nonconvex objectives. Existing accelerated BPG algorithms for convex objectives typically require additional assumptions on Bregman divergences, most notably triangle-scaling conditions, which can lead to slower convergence rates. We propose a family of geometry-accelerated BPG algorithms that exploit Bregman geometry in both proximal-gradient and mirror-space updates, without imposing additional geometric conditions such as triangle-scaling conditions. Our methods adapt the stepsizes and mirror-space updates through local backtracking and computable acceptance criteria, without a global relative-smoothness constant as input. We derive convergence bounds in terms of the parameters accepted during the iterations. These bounds yield an $\mathcal{O}(k^{-2})$ rate for convex objectives and a linear rate for relatively strongly convex objectives when the geometry-acceleration parameters remain uniformly bounded. For nonconvex objectives, we establish an $\mathcal{O}(k^{-1})$ rate for a stationarity measure without requiring a lower Bregman bound or a full-domain Bregman divergence. Numerical experiments on inverse problems, entropy-regularized least squares, D-optimal design, and nonnegative matrix factorization demonstrate faster practical convergence than established Bregman baselines.

1 INTRODUCTION

We focus on minimizing the sum of a smooth function and a nonsmooth convex function:

$$\min_{x \in \operatorname{cl}\Omega} \quad \Phi(x) \coloneqq f(x) + \rho(x), \tag{1}$$

where $\operatorname{cl}\Omega$ is the closure of a nonempty open convex set $\Omega \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, lower semicontinuous, and continuously differentiable on Ω, and $\rho : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, lower semicontinuous, and convex. Relative smoothness (Bauschke et al., 2017; Bolte et al., 2018; Lu et al., 2018) replaces the quadratic upper approximation of f with an upper approximation based on a Bregman divergence. Under this assumption, we study convex, relatively strongly convex, and nonconvex f.

Related work. Classical first-order methods, including proximal gradient methods, are typically analyzed under Euclidean smoothness (see (Beck, 2017; Nesterov, 2018)), *i.e.*, ∇f is Lipschitz continuous. This assumption provides a simple quadratic upper bound on the objective and enables accelerated algorithms with fast convergence rates (Nesterov, 1983; Beck & Teboulle, 2009). However, some objectives fall outside this setting because their curvature is unbounded over the effective domain, for example near its boundary. Bregman geometry, which also underlies mirror descent (Nemirovski & Yudin, 1983), allows upper bounds better adapted to objective curvature, providing a basis for Bregman proximal gradient (BPG) algorithms (Bauschke et al., 2017; Bolte et al., 2018; Lu et al., 2018). BPG algorithms and their variants yield convergence guarantees without global Euclidean smoothness, with applications to Poisson linear inverse problems (Bauschke et al., 2017) and nonconvex inverse problems such as phase retrieval (Bolte et al., 2018; Takahashi et al., 2022), blind deconvolution (Takahashi et al., 2023), and nonnegative matrix factorization (Mukkamala & Ochs, 2019; Takahashi et al., 2026b). See also Appendix A for other methods based on Bregman divergences.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Acceleration of BPG. In Euclidean geometry, accelerated proximal gradient algorithms attain the $\mathcal{O}(k^{-2})$ convergence rate after k iterations for convex objectives. Extending such acceleration to Bregman geometry is substantially more difficult. In fact, under relative smoothness alone, the $\mathcal{O}(k^{-1})$ rate is optimal for a broad class of Bregman first-order algorithms (Dragomir et al., 2022). Therefore, achieving a faster global rate for this broad class requires additional structure beyond relative smoothness. Hanzely et al. (2021) obtain an $\mathcal{O}(k^{-\gamma})$ rate by imposing a triangle-scaling inequality on the Bregman divergence, where the triangle-scaling exponent $\gamma \in (0, 2]$ controls how the divergence scales under linear coupling. Many subsequent accelerated Bregman algorithms have continued to rely on triangle-scaling conditions (Liu & Liu, 2022; Savchuk et al., 2025). On the other hand, the triangle-scaling exponent can deteriorate substantially for some Bregman divergences; in particular, Hanzely et al. (2021) report that no uniform exponent $\gamma > 1/2$ is available for the Itakura–Saito divergence. Hanzely et al. (2021) also develop ABPG-g, which adapts gains using local triangle-scaling tests and bounds the objective gap in terms of the intrinsic exponent γ and the accepted gains. Then, the following question naturally arises:

> *Can we establish accelerated convergence rates under conditions on the parameters chosen during the iterations, without imposing a global triangle-scaling condition on the Bregman divergence?*

Contributions. We answer this question affirmatively by introducing geometry-accelerated BPG (GA-BPG). For convex objectives, its acceptance test controls the auxiliary mirror displacement instead of triangle scaling. We develop three variants, GA-BPGc, GA-BPGsc, and GA-BPGnc, for convex, relatively strongly convex, and nonconvex objectives, respectively. Our main contributions are summarized as follows:

- **Mirror geometry.** GA-BPG constructs auxiliary points in mirror space to guide subsequent BPG steps. Computable criteria determine whether to accept these updates, adjust their parameters, or reset the auxiliary points.
- **Acceleration.** We derive convergence bounds in terms of the parameters accepted during the iterations. For convex and relatively strongly convex objectives, these bounds yield $\mathcal{O}(k^{-2})$ and linear convergence, respectively, when the geometry-acceleration parameters remain uniformly bounded. For nonconvex objectives, we obtain an $\mathcal{O}(k^{-1})$ bound on a stationarity residual. These results recover the corresponding optimal rate orders in Euclidean geometry.
- **Minimal assumptions.** Under the standard BPG assumptions for each regime, we derive convergence bounds that depend on the iterates and parameters generated by the algorithms. GA-BPGc requires neither a triangle-scaling condition nor global curvature constants as input. GA-BPGnc requires neither a lower Bregman bound nor the assumption $\Omega = \mathbb{R}^n$, relying only on upper relative smoothness and lower boundedness of the objective. None of the three algorithms requires a global relative-smoothness constant L as input; GA-BPGsc uses the known relative strong-convexity constant μ.
- **Numerical evaluation.** Experiments on inverse problems, entropy-regularized least squares, D-optimal design, and nonnegative matrix factorization demonstrate faster practical convergence than established Bregman baselines.

Table 1 compares our convergence guarantees with existing results. GA-BPGc and GA-BPGsc attain accelerated rates if the accepted geometry-acceleration parameters are uniformly bounded.

Notation. For vectors, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the Euclidean inner product and norm, respectively. For a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_{\mathrm{F}}$ denotes the Frobenius norm. $\mathbf{1}$ denotes the all-ones vector of appropriate dimension. Vector and matrix inequalities using $\geq$ or $\leq$ are interpreted componentwise. For a set $S \subset \mathbb{R}^n$, $\operatorname{cl} S$ and $\operatorname{int} S$ denote its closure and interior. The indicator function δ_S equals 0 on S and $+\infty$ elsewhere. For an extended-real-valued function $\psi : \mathbb{R}^n \to [-\infty, +\infty]$, its *effective domain* is $\operatorname{dom}\psi := \{x \in \mathbb{R}^n : \psi(x) < +\infty\}$, and its *convex conjugate* is $\psi^*(s) := \sup_{x\in\mathbb{R}^n}\{\langle s, x\rangle - \psi(x)\}$. For convex ψ, $\partial\psi(x)$ denotes its subdifferential, whose elements are *subgradients*. The function ψ is proper if $\operatorname{dom}\psi \neq \emptyset$ and $\psi(x) > -\infty$ for all $x \in \mathbb{R}^n$, and a proper lower semicontinuous convex function is *Legendre* if it is essentially smooth and essentially strictly convex. For a Legendre function ψ, the *inverse mirror map* is $\nabla\psi^* = (\nabla\psi)^{-1}$, with $\nabla\psi^*(s) = \operatorname{argmin}_{x\in\mathbb{R}^n}\{\psi(x) - \langle s, x\rangle\}$

Table 1: **Assumptions and convergence guarantees.** "Cond.": additional conditions; TS_γ: local triangle-scaling with gains and intrinsic exponent $\gamma \in (0, 2]$. "No L input": the global relative-smoothness constant need not be supplied. Rates are for objective gaps in convex (cvx) and relatively strongly convex (r-scvx) problems, and stationarity measures in nonconvex (ncvx) problems. †: rates hold with uniformly bounded geometry-acceleration parameters. ♭: the bound depends on accepted gains; C^2 Bregman generators allow $\gamma = 2$, yielding $\mathcal{O}(k^{-2})$ with bounded geometric-mean gains.

Method	Reference	f	Cond.	No L input	Rate
BPG	Bauschke et al. (2017)	cvx	-	✗	$\mathcal{O}(k^{-1})$
ABPG-g	Hanzely et al. (2021)	cvx	TS_γ	✗	$\mathcal{O}(k^{-\gamma})^\flat$
GA-BPGc	Algorithm 2	cvx	-	✓	$\mathcal{O}(k^{-2})^\dagger$
BPG	Bauschke et al. (2019)	r-scvx	-	✗	linear
GA-BPGsc	Algorithm 3	r-scvx	-	✓	linear†
BPG	Bolte et al. (2018)	ncvx	-	✗	$\mathcal{O}(k^{-1})$
BPGe	Zhang et al. (2019)	ncvx	$\Omega = \mathbb{R}^n$	✗	$\mathcal{O}(k^{-1})$
GA-BPGnc	Algorithm 4	ncvx	-	✓	$\mathcal{O}(k^{-1})$

for $s \in \operatorname{int}\operatorname{dom}\psi^*$. For $u \in \operatorname{dom}\psi$ and $v \in \operatorname{int}\operatorname{dom}\psi$, define the *Bregman divergence* $D_\psi(u, v) \coloneqq \psi(u) - \psi(v) - \langle \nabla\psi(v), u - v\rangle$. We use the same notation throughout, including for nonconvex f. For each $v \in \operatorname{int}\operatorname{dom}\psi$, we extend $D_\psi(u, v)$ to $u \in \operatorname{cl}\operatorname{dom}\psi$ by the same formula, allowing the value $+\infty$. Finally, for $x \in \Omega$, define $\partial\Phi(x) \coloneqq \nabla f(x) + \partial\rho(x)$, and call x stationary if $0 \in \partial\Phi(x)$.

2 Proposed Algorithm: Geometry-Accelerated BPG

We impose the standing assumptions stated below.

Assumption 2.1 (Standing assumptions)**.** The following conditions hold.

(i) $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ is a Legendre function with $\Omega = \operatorname{int}\operatorname{dom}\psi$.

(ii) $f : \mathbb{R}^n \to (-\infty, +\infty]$ is proper and lower semicontinuous with $\operatorname{dom}\psi \subset \operatorname{dom} f$, and is continuously differentiable on Ω.

(iii) $\rho : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, lower semicontinuous, and convex with $\Omega \cap \operatorname{dom}\rho \neq \emptyset$.

(iv) **Relative smoothness.** (Bauschke et al., 2017; Lu et al., 2018) f is *relatively smooth* with respect to ψ, *i.e.*, there exists $L > 0$ such that $D_f(u, v) \leq L D_\psi(u, v)$ holds for all $u, v \in \Omega$.

(v) The objective is bounded below, *i.e.*, $\Phi^\star \coloneqq \inf_{x \in \operatorname{cl}\Omega} \Phi(x) > -\infty$.

Assumption 2.1 is standard in first-order optimization based on Bregman divergences (Bolte et al., 2018; Hanzely et al., 2021; Takahashi et al., 2022; 2026b). Relative smoothness in Assumption 2.1(iv) ensures convergence of BPG algorithms and is also called the smooth adaptable property (Bolte et al., 2018). Examples of commonly used BPG geometries ψ are listed in Table 2.

Geometry-accelerated BPG. Gradient descent and proximal gradient descent use squared Euclidean distances to construct subproblems, whereas mirror descent and BPG use Bregman divergences. However, BPG can converge slowly on (1) in practice. Dragomir et al. (2022) prove $\mathcal{O}(1/k)$ worst-case optimal for their Bregman first-order methods under convexity and relative smoothness alone. Hanzely et al. (2021) obtain an $\mathcal{O}(1/k^\gamma)$ rate under a uniform triangle-scaling condition with exponent $\gamma \in (0, 2]$. The exponent depends on the Bregman divergence and can be far below 2, giving slower convergence rates. For the Itakura–Saito divergence, no positive uniform exponent exists on its effective domain. ABPG-g adapts gains using local triangle-scaling tests, with a bound determined by the intrinsic exponent γ and the accepted gains (Hanzely et al., 2021). Our convex methods use acceptance tests on auxiliary mirror displacements to obtain accelerated bounds.

We now present the common template for the three *geometry-accelerated BPG algorithms* (GA-BPG) in Algorithm 1: GA-BPGc (Algorithm 2) for convex objectives, GA-BPGsc (Algorithm 3) for

Algorithm 1: Common GA-BPG template

Input: A Legendre function ϕ; $y_0 \in \Omega \cap \operatorname{dom}\rho \cap \operatorname{int}\operatorname{dom}\phi$; $\lambda_0 > 0$; $\gamma_+ \geq 1$, $\gamma_- > 1$.

$z_0 \leftarrow y_0$

for $k = 1, 2, \ldots$ **do**

  Initialize $\lambda \leftarrow \gamma_+ \lambda_{k-1}$.

  Coupling. Form the coupled iterate x using y_{k-1} and z_{k-1}.

  BPG update. Compute

$$\widehat{y} \in T_\lambda(x) := \operatorname*{argmin}_{u \in \operatorname{cl}\Omega} \left\{ \rho(u) + \langle \nabla f(x), u - x \rangle + \frac{1}{\lambda} D_\psi(u, x) \right\}, \tag{2}$$

  if $D_f(\widehat{y}, x) > \lambda^{-1} D_\psi(\widehat{y}, x)$ **then**

    $\lambda \leftarrow \lambda / \gamma_-$ and retry.

  Mirror update. Construct an auxiliary point $\widehat{z}$ via an inverse mirror map $\nabla\phi^*$.

  Criterion. Select y_+ from $\widehat{y}$ or y_{k-1}, and determine whether to accept, retry, or reset $\widehat{z}$.

  $(x_k, y_k, z_k, \lambda_k) \leftarrow (x, y_+, \widehat{z}, \lambda)$

relatively strongly convex objectives, and GA-BPGnc (Algorithm 4) for nonconvex objectives. Each consists of four components: (i) coupling, (ii) BPG update, (iii) criterion, and (iv) mirror update, whose order may vary. We explain this structure through GA-BPGc; GA-BPGsc and GA-BPGnc are discussed separately in Sections 3.2 and 3.3, respectively.

Coupling. Let $\lambda > 0$ denote the BPG stepsize in (2) and let $\kappa > 0$ denote the geometry-acceleration parameter. GA-BPGc and GA-BPGsc combine y_{k-1} with the auxiliary point z_{k-1}:

$$\omega = \omega_{k-1} + \alpha, \qquad x = \frac{\omega_{k-1} y_{k-1} + \alpha z_{k-1}}{\omega},$$

where ω_k is the cumulative weight ($\omega_0 = 0$). GA-BPGc backtracks κ until acceptance, adjusting the coupling weight and mirror update scale via $\kappa\alpha^2 = 2\lambda\omega$ ($\alpha > 0$).

BPG update. All GA-BPG algorithms compute $\widehat{y} \in T_\lambda(x)$ by solving (2), where x is the point obtained from the coupling step. We make the following assumption for the subproblem.

Assumption 2.2. For every $x \in \Omega$ and $\lambda > 0$, subproblem (2) has a minimizer.

Assumption 2.2 holds, for example, if $\psi + \lambda\rho$ is supercoercive for every $\lambda > 0$ (Bauschke et al., 2017, Lemma 2(ii)). The solution set $T_\lambda(x)$ is a singleton contained in Ω; see Proposition B.1. All GA-BPG algorithms also use backtracking until $\widehat{y}$ satisfies the relative smoothness inequality

$$D_f(\widehat{y}, x) \leq \lambda^{-1} D_\psi(\widehat{y}, x). \tag{3}$$

The complexity of backtracking is discussed in Appendix B.4.

Mirror update. Let $g_\lambda(x) := (\nabla\psi(x) - \nabla\psi(\widehat{y}))/\lambda$, where $\widehat{y} \in T_\lambda(x)$. GA-BPGc uses a Legendre function $\phi : \mathbb{R}^n \to (-\infty, +\infty]$, possibly different from ψ, and sets

$$\widehat{z} = \nabla\phi^*(\nabla\phi(z_{k-1}) - \alpha g_\lambda(x)).$$

Practical guidance on choosing the auxiliary geometry ϕ, together with representative pairs of BPG and auxiliary geometries (ψ, ϕ), is provided by Table 2 in Appendix C.1.

Criterion. Unlike the other three components, this component handles auxiliary calculations, domain checks, and acceptance decisions at multiple points in each algorithm. For GA-BPGc and GA-BPGsc, the acceptance inequality keeps the Lyapunov function nonincreasing. A failed mirror-domain or acceptance check increases κ and triggers a new trial; see Section 3 for details. On acceptance, both set $y_k = \widehat{y}$ and record the accepted geometry-acceleration parameter $\kappa_k = \kappa$.

Termination. Stopping tests are omitted from the pseudocode. GA-BPGc and GA-BPGsc return y_{k-1} before the inner search if $y_{k-1} \in T_{\lambda_{k-1}}(y_{k-1})$. All three algorithms return x whenever a BPG trial gives $\widehat{y} = x$. The numerical stopping rules used in the experiments are described in Section 4.

Algorithm 2: GA-BPGc

Input: A Legendre function ϕ; $y_0 \in \Omega \cap \operatorname{dom}\rho \cap \operatorname{int}\operatorname{dom}\phi$; $\lambda_0, \kappa_0 > 0$; $\gamma_+ \geq 1$; $\gamma_-, \gamma_\kappa > 1$.

$z_0 \leftarrow y_0, \omega_0 \leftarrow 0$

for $k = 1, 2, \ldots$ **do**

  $(\widehat{\lambda}, \widehat{\kappa}) \leftarrow (\gamma_+\lambda_{k-1}, \kappa_{k-1}/\gamma_+)$

  $(\alpha_k, \omega_k, x_k, y_k, z_k, \lambda_k, \kappa_k) \leftarrow \texttt{Backtracking}\,(\omega_{k-1}, y_{k-1}, z_{k-1}, \widehat{\lambda}, \widehat{\kappa})$

Procedure `Backtracking` $(\omega_-, y, z, \lambda, \kappa)$

  $\alpha \leftarrow (\lambda + \sqrt{\lambda^2 + 2\kappa\lambda\omega_-})/\kappa$, $\omega \leftarrow \omega_- + \alpha$, $x \leftarrow (\omega_- y + \alpha z)/\omega$ ▷ `Coupling`

  $\widehat{y} \in T_\lambda(x)$ ▷ `BPG update`

  if $D_f(\widehat{y}, x) > D_\psi(\widehat{y}, x)/\lambda$ **then**

    return `Backtracking` $(\omega_-, y, z, \lambda/\gamma_-, \kappa)$

  $p \leftarrow \nabla\phi(z) - \alpha(\nabla\psi(x) - \nabla\psi(\widehat{y}))/\lambda$

  if $p \notin \operatorname{int}\operatorname{dom}\phi^*$ **then**

    return `Backtracking` $(\omega_-, y, z, \lambda, \gamma_\kappa\kappa)$

  $\widehat{z} \leftarrow \nabla\phi^*(p)$ ▷ `Mirror update`

  if $\widehat{z} \notin \Omega$ ***or*** $D_\phi(z, \widehat{z}) > \omega P_\lambda(x) + \omega_- D_f(y, x)$ **then** ▷ `Criterion`

    return `Backtracking` $(\omega_-, y, z, \lambda, \gamma_\kappa\kappa)$

  return $(\alpha, \omega, x, \widehat{y}, \widehat{z}, \lambda, \kappa)$

3 Convergence analysis

This section gives complexity bounds for all three algorithms. For the unique $\widehat{y} \in T_\lambda(x)$, define

$$P_\lambda(x) := \frac{D_\psi(x, \widehat{y}) + D_\psi(\widehat{y}, x)}{\lambda} - D_f(\widehat{y}, x).$$

The following comparison inequality explains the role of P_λ. Its proof is given in Appendix B.2.

Lemma 3.1 (BPG comparison)**.** *Suppose that Assumptions 2.1 and 2.2 hold. Let $\widehat{y} \in T_\lambda(x)$, with $x \in \Omega$ and $\lambda > 0$. For every $u \in \operatorname{cl}\Omega \cap \operatorname{dom}\Phi$,*

$$\Phi(\widehat{y}) - \Phi(u) \leq \langle g_\lambda(x), x - u\rangle - P_\lambda(x) - D_f(u, x). \tag{4}$$

If (3) holds, then $P_\lambda(x) \geq \lambda^{-1} D_\psi(x, \widehat{y}) \geq 0$, with equality $P_\lambda(x) = 0$ if and only if $x = \widehat{y}$.

In Sections 3.1 and 3.2, u is a comparison point used only in the analysis. Taking $u = x^\star$ yields optimality-gap bounds whenever $x^\star$ lies in the comparison set.

3.1 Convex objectives

We have a complexity bound for GA-BPGc for convex objective functions.

Assumption 3.2 (Convexity)**.** The function f is convex.

Take $u \in \mathcal{U}_\phi$, where $\mathcal{U}_\phi := \operatorname{cl}\Omega \cap \operatorname{dom}\Phi \cap \operatorname{dom}\phi$, and define the Lyapunov function as follows:

$$\mathcal{E}_k^{\mathrm{C}}(u) := \omega_k(\Phi(y_k) - \Phi(u)) + D_\phi(u, z_k).$$

Combining two applications of Lemma 3.1 with the coupling and mirror identities in Appendix B.5 yields the following estimate; see Appendix C.2 for the proof.

Lemma 3.3 (Convex one-step estimate)**.** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold. For every accepted iteration $k \geq 1$ of GA-BPGc and every $u \in \mathcal{U}_\phi$,*

$$\mathcal{E}_k^{\mathrm{C}}(u) - \mathcal{E}_{k-1}^{\mathrm{C}}(u) \leq D_\phi(z_{k-1}, z_k) - \omega_k P_{\lambda_k}(x_k) - \omega_{k-1} D_f(y_{k-1}, x_k) - \alpha_k D_f(u, x_k). \tag{5}$$

Stopping checks ensure finitely many BPG solves per iteration (Appendix C.4). The acceptance test

$$D_\phi(z_{k-1}, \widehat{z}) \leq \omega P_\lambda(x) + \omega_{k-1} D_f(y_{k-1}, x) \tag{6}$$

uses only the current trial, not u, and makes the sum of the first three terms on the right of (5) nonpositive. Since $D_f(u, x_k) \geq 0$ by Assumption 3.2, every accepted iteration satisfies $\mathcal{E}_k^{\mathrm{C}}(u) \leq \mathcal{E}_{k-1}^{\mathrm{C}}(u)$. The coefficient equation gives the following weight-growth bound, converting this decrease into an objective bound; see Appendix C.3 for the proof.

Lemma 3.4 (Growth of the convex weights)**.** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold. For the accepted parameters of GA-BPGc,* $\omega_N \geq \frac{1}{2}\left(\sum_{k=1}^{N}\sqrt{\lambda_k/\kappa_k}\right)^2$.

Using Lemma 3.4, we obtain the following convergence bound in terms of the accepted stepsizes λ_k and geometry-acceleration parameters κ_k; its proof is given in Appendix C.3.

Theorem 3.5 (Convergence bound for convex objectives)**.** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold. After* $N \geq 1$ *completed iterations of GA-BPGc, for every* $u \in \mathcal{U}_\phi$,

$$\Phi(y_N) - \Phi(u) \leq \frac{2D_\phi(u, z_0)}{\left(\sum_{k=1}^{N}\sqrt{\lambda_k/\kappa_k}\right)^2}. \tag{7}$$

If a minimizer $x^\star \in \operatorname{cl}\Omega$ *belongs to* $\operatorname{dom}\phi$*, this is an optimality-gap bound with* $u = x^\star$*. In particular,* $\lambda_k/\kappa_k \geq c > 0$ *yields* $\Phi(y_N) - \Phi^\star = \mathcal{O}(N^{-2})$*. For Burg–entropy geometry, bounded sublevel sets and an interior minimizer ensure* $\sup_k \kappa_k < \infty$ *and* $\mathcal{O}(k^{-2})$ *convergence without triangle scaling (Appendix C.5), while Appendix C.6 recovers the Euclidean accelerated rate.*

3.2 Relatively strongly convex objectives

We next consider GA-BPGsc in the relatively strongly convex setting.

Assumption 3.6 (Relative strong convexity)**.** f is relatively strongly convex with respect to ψ, *i.e.*, there exists a constant $\mu > 0$ such that $D_f(u, v) \geq \mu D_\psi(u, v)$ holds for all $u \in \operatorname{cl}\Omega$ and $v \in \Omega$.

Take $u \in \mathcal{U}_\psi$, where $\mathcal{U}_\psi \coloneqq \operatorname{cl}\Omega \cap \operatorname{dom}\Phi \cap \operatorname{dom}\psi$, and define the Lyapunov function as follows:

$$\mathcal{E}_k^{\mathrm{S}}(u) \coloneqq \omega_k(\Phi(y_k) - \Phi(u)) + \theta_k D_\psi(u, z_k). \tag{8}$$

We present GA-BPGsc in Algorithm 3 in Appendix D and explain its structure here.

- **Coupling.** Let $\alpha > 0$ satisfy $\kappa\alpha^2 = 2\lambda\omega\theta$, where $\theta = \theta_{k-1} + \mu\alpha$ and $\kappa > 2\mu\lambda$. The weights are initialized with $\omega_0 = 0$, and $\theta_0 = 1$.
- **Mirror update.** Set $\phi = \psi$ in GA-BPGc and replace its mirror update with
$$\widehat{z} = \nabla\psi^*\left(\frac{\theta_{k-1}\nabla\psi(z_{k-1}) + \mu\alpha\nabla\psi(x) - \alpha g_\lambda(x)}{\theta}\right).$$
- **Criterion.** GA-BPGsc accepts the trial if
$$\theta D_\psi(z_{k-1}, \widehat{z}) - \mu\left(\omega_{k-1}D_\psi(y_{k-1}, x) + \alpha D_\psi(z_{k-1}, x)\right) \leq \omega P_\lambda(x). \tag{9}$$
By Lemma 3.7, this ensures that the Lyapunov function $\mathcal{E}_k^{\mathrm{S}}(u)$ does not increase. With the stopping checks, each iteration finishes after finitely many BPG solves (Appendix D.3).

The coupling and weighted mirror identities in Appendix B.5, together with Lemma 3.1 and relative strong convexity, yield the following estimate. See Appendix D.2 for the proof.

Lemma 3.7 (Relatively strongly convex one-step estimate)**.** *Suppose that Assumptions 2.1, 2.2 and 3.6 hold. For every accepted iteration* $k \geq 1$ *of GA-BPGsc and every* $u \in \mathcal{U}_\psi$,

$$\begin{aligned}\mathcal{E}_k^{\mathrm{S}}(u) - \mathcal{E}_{k-1}^{\mathrm{S}}(u) \leq{}& \theta_k D_\psi(z_{k-1}, z_k) - \omega_k P_{\lambda_k}(x_k) \\ &- \mu\left(\omega_{k-1}D_\psi(y_{k-1}, x_k) + \alpha_k D_\psi(z_{k-1}, x_k)\right).\end{aligned} \tag{10}$$

The criterion (9) uses only the current trial, not u, and ensures $\mathcal{E}_k^{\mathrm{S}}(u) \leq \mathcal{E}_{k-1}^{\mathrm{S}}(u)$. The coefficient equation links this decrease to an objective bound. For accepted parameters, $\kappa_k > 2\mu\lambda_k$ ensures $\sqrt{2\mu\lambda_k/\kappa_k} \in (0, 1)$. The weights satisfy

$$\omega_N \geq \omega_1 \prod_{k=2}^{N}\left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right)^{-1}, \qquad N \geq 1, \tag{11}$$

where an empty product equals one. Using this growth estimate, we obtain the following convergence bound in terms of the accepted ratios λ_k/κ_k; the proofs of the growth estimate and the theorem are given in Appendix D.2.

Theorem 3.8 (Convergence bound for relatively strongly convex objectives)**.** *Suppose that Assumptions 2.1, 2.2 and 3.6 hold. After $N \geq 1$ completed iterations of GA-BPGsc, for every $u \in \mathcal{U}_\psi$,*

$$\Phi(y_N) - \Phi(u) \leq \frac{D_\psi(u, z_0)}{\omega_1} \prod_{k=2}^{N} \left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right). \tag{12}$$

If a minimizer $x^\star \in \operatorname{cl}\Omega$ belongs to $\operatorname{dom}\psi$, this is an optimality-gap bound with $u = x^\star$. In particular, a uniform bound $\sqrt{2\mu\lambda_k/\kappa_k} \geq q_{\min} > 0$ yields linear convergence of the objective gap. For Burg geometry, the stated assumptions already ensure $\sup_k \kappa_k < \infty$ and linear convergence (Appendix D.4), while Appendix D.5 recovers the Euclidean accelerated rate.

3.3 Nonconvex objectives

For $x \in \Omega \cap \operatorname{dom}\rho$ and $\widehat{y} \in T_\lambda(x)$, define

$$\mathcal{R}_\lambda(x) := -\frac{2}{\lambda}\left(\langle \nabla f(x), \widehat{y} - x\rangle + \rho(\widehat{y}) - \rho(x) + \lambda^{-1} D_\psi(\widehat{y}, x)\right).$$

This normalized Bregman forward–backward residual vanishes exactly at interior stationary points and satisfies $\mathcal{R}_\lambda(x) \geq 2D_\psi(x, \widehat{y})/\lambda^2$; see Appendix B.3. The envelope interpretation is discussed in Ahookhosh et al. (2021); Fatkhullin & He (2024). We present GA-BPGnc in Algorithm 4 in Appendix E and explain its structure here.

- **Coupling & mirror update.** GA-BPGnc couples y_k and y_{k-1} by extrapolation in the mirror space induced by a fixed Legendre function ϕ, with an arbitrary weight $\beta_{k+1} \geq 0$:
$$\widehat{z} = \nabla\phi^*(\nabla\phi(y_k) + \beta_{k+1}(\nabla\phi(y_k) - \nabla\phi(y_{k-1}))).$$
We set $z_k = \widehat{z}$ if extrapolation is permitted and $\widehat{z}$ is well defined with $\widehat{z} \in \Omega \cap \operatorname{dom}\rho$; otherwise, we set $z_k = y_k$ and start the next BPG update from y_k.
- **Criterion.** GA-BPGnc sets $y_k = \widehat{y}$ if $\Phi(\widehat{y}) < \Phi(y_{k-1})$, and $y_k = y_{k-1}$ otherwise. GA-BPGnc rejects the extrapolation if $\delta_k < \frac{\sigma\lambda}{2}\mathcal{R}_\lambda(x)$, where $\delta_k := \Phi(y_{k-1}) - \Phi(\widehat{y})$ and $\sigma \in (0, 1)$ is the reset parameter.

Take a nonterminal iteration, for which $\mathcal{R}_{\lambda_k}(x_k) > 0$, and define $[t]_+ := \max(t, 0)$ for $t \in \mathbb{R}$ and

$$\tau_k := \frac{2\delta_k}{\lambda_k \mathcal{R}_{\lambda_k}(x_k)}.$$

The acceptance rule relates the objective decrease to $\frac{\lambda_k}{2}\mathcal{R}_{\lambda_k}(x_k)$; the reset rule then ensures $\tau_{k+1} \geq 1$ whenever $\tau_k < \sigma$ and the next iteration is nonterminal. Its proof is given in Section E.2.

Lemma 3.9 (Descent and a safe next iteration)**.** *Suppose that Assumptions 2.1 and 2.2 hold. Every nonterminal iteration of GA-BPGnc satisfies*

$$\Phi(y_{k-1}) - \Phi(y_k) = [\delta_k]_+ = \frac{\lambda_k}{2}[\tau_k]_+ \mathcal{R}_{\lambda_k}(x_k). \tag{13}$$

If $x_k = y_{k-1}$, then $\tau_k \geq 1$; if $\tau_k < \sigma$, then $x_{k+1} = y_k$; hence $\tau_{k+1} \geq 1$ if step $k+1$ is nonterminal.

Since $\frac{\lambda_k}{2}[\tau_k]_+ \mathcal{R}_{\lambda_k}(x_k) \geq 0$, the acceptance rule ensures $\Phi(y_k) \leq \Phi(y_{k-1})$. Summing these decreases gives a stationarity bound in terms of $\sum_{k\in\mathcal{A}_N} \tau_k\lambda_k$, with $\mathcal{A}_N$ defined below. The reset rule and stepsize lower bound control this sum from below, yielding an $\mathcal{O}(N^{-1})$ worst-case bound. The following theorem summarizes these bounds; see Appendices E.3, E.4, and E.5 for the proof.

Theorem 3.10 (Stationarity bounds for nonconvex objectives)**.** *Suppose that Assumptions 2.1 and 2.2 hold and GA-BPGnc has not returned a stationary point in its first $N \geq 2$ iterations. Let $\mathcal{A}_N := \{1 \leq k \leq N : \delta_k > 0\}$ and $\underline{\lambda} := \min\{\lambda_0, 1/(\gamma_- L)\} > 0$. Then*

$$\min_{k\in\mathcal{A}_N} \mathcal{R}_{\lambda_k}(x_k) \leq \frac{2(\Phi(y_0) - \Phi^\star)}{\sum_{k\in\mathcal{A}_N} \tau_k\lambda_k} \leq \frac{4(\Phi(y_0) - \Phi^\star)}{\sigma\underline{\lambda}(N-1)}. \tag{14}$$

Thus a residual at most ε^2 is obtained within $\mathcal{O}(\varepsilon^{-2})$ outer iterations and BPG subproblem solves.

The first bound in (14) depends on the values of τ_k and λ_k obtained during the iterations. The second follows from the reset rule and the stepsize lower bound; it depends only on N, the initial objective gap, σ, and $\underline{\lambda}$. Unlike Sections 3.1 and 3.2, the proof sums objective decreases rather than using comparison-point Lyapunov bounds. Under a uniform subgradient comparison, the residual bound implies a norm-based stationarity bound (Appendix E.6).

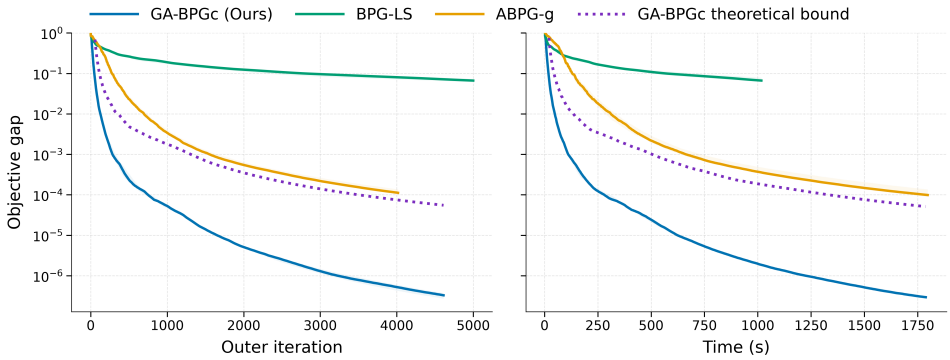

Figure 1: Log plot of the normalized objective gap for convex Poisson on the NYTimes data.

4 Numerical Experiments

We compare GA-BPG variants with Bregman baselines using objective gaps versus iterations and wall-clock time. For an objective Φ and reference value Φ_{ref}, we plot the normalized objective gap $(\Phi(x_k) - \Phi_{\mathrm{ref}})/(\Phi(x_0) - \Phi_{\mathrm{ref}})$, where x_k is the reported iterate. The reference is a known optimal value, an empirical objective value, or a certified lower bound, as indicated below. GA-BPGc and GA-BPGsc use $\gamma_+ = 1/0.9$, $\gamma_- = 2$, and $\gamma_\kappa = 1.5$. GA-BPGnc uses $\gamma_+ = 1.1$, $\gamma_- = 2$, $\sigma = 0.5$, and $\lambda_{\max} = 10^6\lambda_0$. Additional experiments and more discussions are given in Appendices F.1 and F.2, respectively. All experiments were run in Python 3.12.3 on an Intel Core i7-13620H under Ubuntu 24.04 (WSL2).

Convex Poisson. Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, we consider the convex Poisson inverse problem $\min_{x\in\mathbb{R}^n} \Phi_0(x) := \frac{1}{m} D_{\mathrm{KL}}(b, Ax) + \delta_{[0,10^3]^n}(x)$, where $D_{\mathrm{KL}}(u, v) = \sum_i (u_i \log \frac{u_i}{v_i} - u_i + v_i)$ is the Kullback–Leibler divergence. We use the NYTimes bag-of-words data (Newman, 2008), which originally contain 300,000 documents and 102,660 terms. After removing empty rows and columns and scaling each retained row to sum to 100, we obtain a sparse matrix A with $m = 299{,}752$ and $n = 101{,}636$. For each seed, we generate a positive planted vector $x^\star \in (0, 10^3)^n$ and set $b = Ax^\star$, *i.e.*, $\Phi_0^\star = \Phi_0(x^\star) = 0$, which we use as the reference value. All methods use the Burg geometry $\psi(x) := -\sum_j \log x_j$ for (2). GA-BPGc additionally uses the negative-entropy auxiliary geometry $\phi(x) = \sum_j (x_j \log x_j - x_j)$ for its mirror update. All methods start from $y_0 = \mathbf{1}$. Runs stop upon meeting algorithm-specific stopping criteria or reaching 5,000 iterations or 1,800 seconds. Figure 1 compares GA-BPGc with BPG using linesearch (BPG-LS) and ABPG-g (Hanzely et al., 2021), with dotted bounds from (24). Our algorithm outperforms the baselines.

Relatively strongly convex objectives. Motivated by stacked regression (Breiman, 1996) and entropy-penalized convex aggregation (Koltchinskii, 2009), we consider least-squares aggregation with a Kullback–Leibler penalty toward uniform weights: $\min_{w\in\Delta_d} F_\mu(w) := \frac{1}{2m}\|Hw - b\|^2 + \mu D_{\mathrm{KL}}(w, \pi)$, where $\pi = \frac{1}{d}\mathbf{1}$ and $\Delta_d = \{w \in \mathbb{R}^d : w \geq 0,\ \mathbf{1}^\top w = 1\}$, $b \in \mathbb{R}^m$ is the response vector, and $H \in \mathbb{R}^{m\times d}$ is a dense out-of-fold prediction matrix constructed from simulated Gaussian regression data using 10-fold cross-validation. The columns of H comprise 1600 backward-selected least-squares predictors and 1600 ridge predictors, giving $m = 2400$ and $d = 3200$. Working on the affine hull of Δ_d, all algorithms use $f = F_\mu$, $\rho = 0$, and $\psi(w) = \sum_j w_j \log w_j$, with $f, \psi = +\infty$ outside Δ_d. The objective is μ-strongly convex relative to ψ; GA-BPGsc uses $\phi = \psi$ with the normalized exponential inverse mirror map. All methods use $w_0 = \pi$ and $\mu = 10^{-3}$. Runs stop upon meeting algorithm-specific criteria or reaching 10,000 iterations or 600 seconds. Figure 2 compares GA-BPGsc with BPG-LS and ABPG-g (Hanzely et al., 2021), with dotted bounds from (26). For each seed, we define F_{ref} as the smallest objective value over all recorded primary iterates of the three methods. GA-BPGsc reaches high-accuracy solutions fastest in wall-clock time.

Nonnegative matrix factorization. We factorize the MovieLens 100K ratings by solving $\min_{W\geq 0,\, H\geq 0} F(W, H) := \frac{1}{2|\mathcal{O}|}\sum_{(i,j)\in\mathcal{O}} (W_{i:}H_{:j} - r_{ij})^2$, where $\mathcal{O}$ contains the 100,000 observed

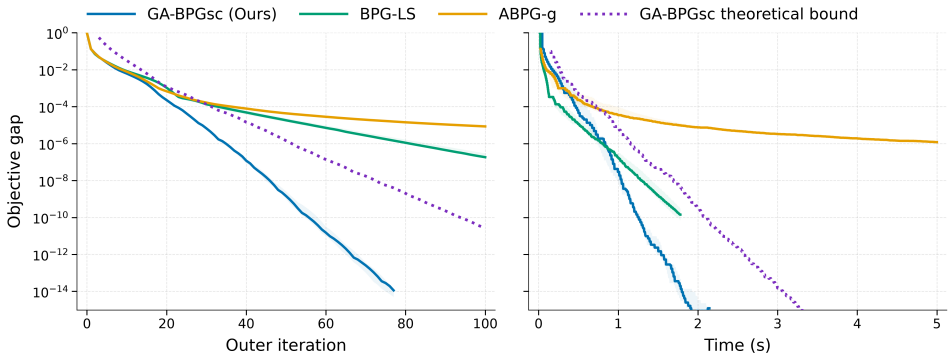

Figure 2: Log plot of the normalized objective gap for entropy-regularized least squares with $\mu = 10^{-3}$.

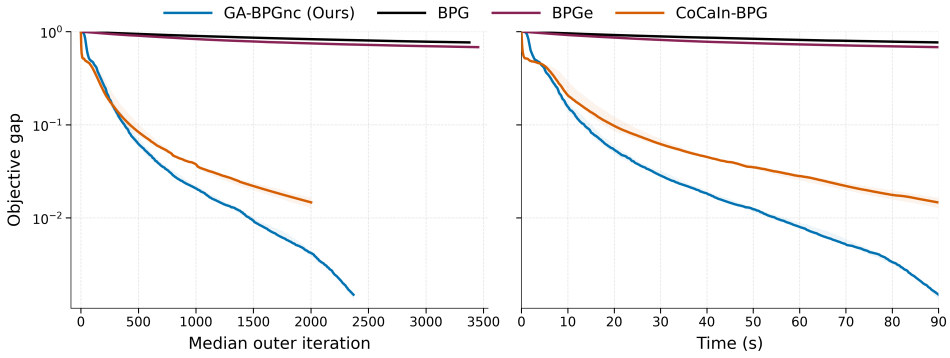

Figure 3: Log plot of the normalized empirical objective gap for nonnegative matrix factorization on MovieLens 100K.

ratings, $W \in \mathbb{R}^{943\times 16}$, and $H \in \mathbb{R}^{16\times 1682}$. Missing ratings are excluded from the objective. We compare GA-BPGnc with BPG (Bolte et al., 2018), BPGe (Zhang et al., 2019), and CoCaIn-BPG (Mukkamala et al., 2020). All methods use $\psi(x) = \frac{1}{4}(\|W\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2)^2 + \frac{1}{2}(\|W\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2)$; GA-BPGnc also uses $\phi = \psi$. BPG and BPGe use the step $1/L$, and GA-BPGnc and CoCaIn-BPG use their respective adaptive rules. The five seeds specify common initializers for all methods. Each method runs for at most 100 seconds. For each seed, F_{ref} is the smallest objective recorded across all four methods during these runs. Figure 3 shows the median and interquartile range of the normalized empirical objective gap over the first 90 seconds. GA-BPGnc reaches smaller empirical gaps than the baselines in both outer iterations and wall-clock time.

5 Conclusion

We proposed GA-BPG, a family of BPG algorithms with auxiliary mirror updates, for convex, relatively strongly convex, and nonconvex objectives. Auxiliary mirror points guide BPG steps, with computable tests for acceptance, parameter updates, and resets. Under standard BPG assumptions, GA-BPGc and GA-BPGsc attain $\mathcal{O}(k^{-2})$ and linear convergence, respectively, when $\sup_k \kappa_k < \infty$. For GA-BPGnc, we established an $\mathcal{O}(k^{-1})$ bound on a stationarity residual without assuming $\Omega = \mathbb{R}^n$. The algorithms require no triangle-scaling condition or prior knowledge of the global relative-smoothness constant. Experiments show faster convergence than the Bregman baselines in all three regimes.

We proved that κ_k remains uniformly bounded for Burg entropy and squared Euclidean distances. For general Bregman divergence, however, κ_k may grow without bound, which can cause the convergence rates to deteriorate. Identifying global conditions that characterize the behavior of κ_k and developing methods to avoid this deterioration remain topics for future work.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing key proof ingredients, assisting with proof writing, proposing or refining hypotheses, providing feedback on research methodology and experiments, implementing algorithms, assisting with translation, developing theoretical models or conceptual frameworks, and cleaning and reformatting datasets. We have not used generative AI tools for directly generating synthetic datasets, or qualitative and thematic data analysis. Additionally, we used generative AI tools for creating or modifying scientific figures, writing or editing code, drafting manuscript text, reviewing our discussion of related work, searching for information, editing for readability, identifying relevant literature. We have reviewed all AI-assisted work. The code was checked by one author, and the proofs were verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns the development and theoretical analysis of optimization algorithms, with numerical experiments on synthetic and existing benchmark datasets. We do not identify any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

The algorithms and assumptions are described in Sections 2 and 3, with proofs provided in Sections B to E. Detailed pseudocode is given in Algorithms 2 to 4. The experimental settings and datasets are described in Section 4. The source code for the numerical experiments is included in the supplementary material.

REFERENCES

Masoud Ahookhosh, Andreas Themelis, and Panagiotis Patrinos. A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: Superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31(1):653–685, 2021.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Heinz H. Bauschke, Jérôme Bolte, Jiawei Chen, Marc Teboulle, and Xianfu Wang. On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 182(3):1068–1087, 2019.

Amir Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Alberto De Marchi, Yura Malitsky, and Adrien B Taylor. Mirror descent algorithms with logarithmic barriers. *arXiv preprint arXiv:2608.22834*, 2026.

Kuangyu Ding and Kim-Chuan Toh. Non-KKT Accumulation in Entropic Mirror Descent. *arXiv preprint arXiv:2608.01658*, 2026.

Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194 (1–2):41–83, 2022.

Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general Bregman divergence. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3493–3501. PMLR, 2024.

Kiwamu Fujiki, Shota Takahashi, and Akiko Takeda. Approximate Bregman proximal gradient algorithm with variable metric Armijo–Wolfe line search. *arXiv preprint arXiv:2510.06615*, 2025.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, 2021.

Jaeyeon Kim, Chanwoo Park, Asuman Ozdaglar, Jelena Diakonikolas, and Ernest K Ryu. Mirror duality in convex optimization. *arXiv preprint arXiv:2311.17296*, 2023.

Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *The Annals of Statistics*, 37(3):1332–1359, 2009.

Frederik Kunstner, Ryan D'Orazio, Victor S Portella, and Adrien Taylor. Mirror Polyak and a Primal-Dual Lifting. *arXiv preprint arXiv:2608.17252*, 2026.

Jin-Zan Liu and Xin-Wei Liu. A dual Bregman proximal gradient method for relatively-strongly convex optimization. *Numerical Algebra, Control and Optimization*, 12(4):679–692, 2022.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Yura Malitsky and Alexander Posch. Entropic mirror descent for linear systems: Polyak's stepsize and implicit bias. *arXiv preprint arXiv:2505.02614*, 2025.

Mahesh Chandra Mukkamala and Peter Ochs. Beyond alternating updates for matrix factorization with inertial Bregman proximal gradient algorithms. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 4268–4278. Curran Associates, Inc., 2019.

Mahesh Chandra Mukkamala, Peter Ochs, Thomas Pock, and Shoham Sabach. Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. *SIAM Journal on Mathematics of Data Science*, 2(3):658–682, 2020.

Arkadii S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, New York, 1983. ISBN 9780471103455.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, Switzerland, 2018.

David Newman. Bag of Words. UCI Machine Learning Repository, 2008.

Hongjia Ou, Puya Latafat, and Andreas Themelis. Linesearch-free adaptive Bregman proximal gradient for convex minimization under local relative smoothness. *Journal of Optimization Theory and Applications*, 210:68, 2026.

Oleg S. Savchuk, Mohammad S. Alkousa, Andrei I. Shushko, Aleksandr A. Vyguzov, Fedor S. Stonyakin, Dmitry A. Pasechnyuk, and Alexander V. Gasnikov. Accelerated Bregman gradient methods for relatively smooth and relatively Lipschitz continuous minimization problems. *Russian Mathematical Surveys*, 80(6):1067–1102, 2025.

Shota Takahashi and Akiko Takeda. Approximate Bregman proximal gradient algorithm for relatively smooth nonconvex optimization. *Computational Optimization and Applications*, 90(1): 227–256, 2025.

Shota Takahashi, Mituhiro Fukuda, and Mirai Tanaka. New Bregman proximal type algorithms for solving DC optimization problems. *Computational Optimization and Applications*, 83:893–931, 2022.

Shota Takahashi, Mirai Tanaka, and Shiro Ikeda. Blind deconvolution with non-smooth regularization via Bregman proximal DCAs. *Signal Processing*, 202:108734, 2023.

Shota Takahashi, Sebastian Pokutta, and Akiko Takeda. Fast Frank–Wolfe Algorithms with Adaptive Bregman Step-Size for Weakly Convex Functions. In *Proceedings of the Fourteenth International Conference on Learning Representations*, pp. 1–38, 2026a.

Shota Takahashi, Mirai Tanaka, and Shiro Ikeda. Majorization-minimization Bregman proximal gradient algorithms for NMF with the Kullback–Leibler divergence. *Journal of Optimization Theory and Applications*, 208(1):1–34, 2026b.

Xiaoya Zhang, Roberto Barrio, María Ángeles Martínez, Hao Jiang, and Lizhi Cheng. Bregman proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *IEEE Access*, 7:126515–126529, 2019.

A Additional related work

Relative smoothness and BPG. Bauschke et al. (2017) replace a Euclidean quadratic upper bound by a Bregman upper bound, allowing proximal-gradient updates beyond globally Lipschitz gradients. Lu et al. (2018) develop relative smoothness and relative strong convexity and analyze first-order algorithms under these conditions. Ou et al. (2026) develop adaptive BPG algorithms for convex composite problems under local relative smoothness, selecting stepsizes from local curvature estimates without backtracking linesearch. Malitsky & Posch (2025) propose Polyak-type stepsizes for entropic mirror descent on consistent linear systems, establishing convergence guarantees and quantifying the implicit bias toward solutions with small ℓ_1 norm when initialized near the origin. Kunstner et al. (2026) analyze mirror Polyak methods based on Bregman projections, with rates adapting to relative smoothness and relative strong convexity, and use a primal–dual lifting to avoid requiring the optimal value for structured convex problems. De Marchi et al. (2026) establish $\mathcal{O}(\log k/k)$ objective convergence rates for mirror descent with logarithmic barriers, including cases with boundary minimizers where standard Bregman-distance bounds are infinite. Beyond proximal methods, Takahashi et al. (2026a) develop Frank–Wolfe algorithms with adaptive Bregman stepsizes for relatively smooth convex and weakly convex objectives.

Acceleration and its limitations. The Euclidean accelerated proximal-gradient rate (Nesterov, 1983; Beck & Teboulle, 2009) does not extend uniformly to arbitrary relatively smooth problems. Dragomir et al. (2022) establish a matching $\mathcal{O}(1/k)$ upper and lower complexity order for the Bregman first-order oracle class considered there. Kim et al. (2023) establish mirror duality between mirror-descent-type methods for reducing objective values and gradient magnitudes, and derive a dual accelerated mirror descent method for convex objectives with Lipschitz gradients and strongly convex distance-generating functions. Hanzely et al. (2021) obtain accelerated rates under triangle-scaling conditions and develop algorithms that adjust gains during the iterations and bound the objective gap in terms of the accepted gains. GA-BPG instead tests the auxiliary mirror displacement through its acceptance inequality. Our bounds yield $\mathcal{O}(k^{-2})$ and linear rates when the accepted ratios λ_k/κ_k are bounded away from zero.

Relative strong convexity and linear convergence. Relative strong convexity supplies a lower bound in the same Bregman geometry as the upper relative-smoothness bound (Lu et al., 2018). Linear convergence beyond the usual Euclidean strong-convexity and Lipschitz-gradient assumptions is also studied by Bauschke et al. (2019), using generalized gradient-domination conditions. Our analysis of GA-BPGsc assumes relative strong convexity and uses the known constant μ in both the weighted mirror update and the acceptance inequality. Its product bound yields a uniform linear rate when the accepted ratios λ_k/κ_k are bounded away from zero.

Nonconvex Bregman algorithms. Bolte et al. (2018) analyze nonconvex BPG algorithms using Bregman upper bounds and smooth-adaptability conditions. BPGe combines BPG with extrapolation (Zhang et al., 2019), whereas CoCaIn-BPG controls extrapolation and stepsizes through upper and lower curvature estimates (Mukkamala et al., 2020). For DC objectives, BPDCA and BPDCAe use the convex components of the objective explicitly (Takahashi et al., 2022). Approximate BPG simplifies subproblems using a local quadratic approximation of the Bregman divergence (Takahashi & Takeda, 2025), and Fujiki et al. (2025) extend this approach with a variable metric linesearch. Ding & Toh (2026) construct relatively smooth nonconvex examples in which bounded entropic mirror-descent sequences have non-KKT boundary accumulation points despite nonincreasing objective values. For GA-BPGnc, we establish a finite-time stationarity bound under the standing assumptions.

Forward–backward envelopes. Bregman forward–backward envelopes provide merit functions and a way to assess candidate directions beyond the underlying BPG update (Ahookhosh et al., 2021). Related normalized envelope gaps are used as stationarity measures in nonconvex mirror-descent analysis (Fatkhullin & He, 2024). We use the gap between the objective and the Bregman forward–backward envelope to define $\mathcal{R}_\lambda$. An additional uniform comparison bound relates this residual to subgradient norms, as stated in Appendix E.6.

B Common BPG properties

Throughout, Assumptions 2.1 and 2.2 hold. BPG base points and solutions belong to Ω. We use the extended first-argument conventions stated in Section 1; no gradient is evaluated at a boundary comparison point. All BPG pairs mentioned below, including limiting pairs and the additional stopping checks, are covered by the solvability assumption.

B.1 The BPG map

Proposition B.1 (Uniqueness and optimality)**.** *For every admissible (x, λ), the minimizer $y \in T_\lambda(x)$ over $\operatorname{cl}\Omega$ is unique, belongs to $\Omega \cap \operatorname{dom}\rho$, and satisfies*

$$g_\lambda(x) - \nabla f(x) \in \partial\rho(y). \tag{15}$$

For $x \in \Omega \cap \operatorname{dom}\rho$,

$$x \in T_\lambda(x) \quad \Longleftrightarrow \quad 0 \in \nabla f(x) + \partial\rho(x). \tag{16}$$

Proof. Fix an admissible pair (x, λ). Existence of a minimizer follows from Assumption 2.2. Set

$$H = \psi + \lambda\rho, \qquad p = \nabla\psi(x) - \lambda\nabla f(x).$$

Since $\operatorname{dom}\psi \subset \operatorname{cl}\Omega$, the BPG subproblem is equivalent to minimizing $H(u) - \langle p, u\rangle$ over $\mathbb{R}^n$.

Let y be any minimizer. The function ψ is continuous at every point of Ω, and $\Omega \cap \operatorname{dom}\rho \neq \emptyset$. Thus the subdifferential sum rule applies, and optimality gives

$$p \in \partial H(y) = \partial\psi(y) + \lambda\partial\rho(y).$$

In particular, $\partial\psi(y) \neq \emptyset$. Essential smoothness of ψ gives $\operatorname{dom}\partial\psi = \operatorname{int}\operatorname{dom}\psi = \Omega$, so $y \in \Omega \cap \operatorname{dom}\rho$. Since $\partial\psi(y) = \{\nabla\psi(y)\}$, the preceding inclusion becomes

$$\frac{\nabla\psi(x) - \nabla\psi(y)}{\lambda} - \nabla f(x) \in \partial\rho(y),$$

which proves (15).

Every minimizer is therefore interior. Strict convexity of ψ on Ω and convexity of ρ make $H(u) - \langle p, u\rangle$ strictly convex on $\Omega \cap \operatorname{dom}\rho$, so the minimizer is unique.

For $x \in \Omega \cap \operatorname{dom}\rho$, the inclusion $x \in T_\lambda(x)$ implies stationarity by (15). Conversely, if $0 \in \nabla f(x) + \partial\rho(x)$, then

$$p \in \nabla\psi(x) + \lambda\partial\rho(x) = \partial H(x).$$

Hence x minimizes the BPG subproblem, so $x \in T_\lambda(x)$. This proves (16). □

Proposition B.2 (Continuity for a fixed stepsize)**.** *Fix $\lambda > 0$. If $x_j \to x \in \Omega$ and the BPG subproblems at x_j and x have interior solutions $y_j \in T_\lambda(x_j)$ and $y \in T_\lambda(x)$, respectively, then $y_j \to y$.*

Proof. Put $H = \psi + \lambda\rho$ and $p(x) = \nabla\psi(x) - \lambda\nabla f(x)$. The function H is proper, lower semicontinuous, and convex, and the BPG subproblem is equivalent to minimizing $H(u) - \langle p(x), u\rangle$ over $\mathbb{R}^n$. Its unique solution implies

$$\partial H^*(p(x)) = T_\lambda(x) = \{y\}.$$

Moreover, $p(x) \in \operatorname{int}\operatorname{dom} H^*$: otherwise a nonzero supporting normal to the convex domain at $p(x)$ could be added to any subgradient, contradicting this singleton identity.

The subdifferential of a finite-dimensional closed convex function is locally bounded in the interior of its domain. Since $p(x_j) \to p(x)$ and $y_j \in \partial H^*(p(x_j))$, these solutions are bounded. Closedness of the subdifferential graph places every cluster point in $\partial H^*(p(x))$. Uniqueness then gives the claimed convergence. □

B.2 Comparison and stopping tests

Proof of Lemma 3.1. Let $y \in T_\lambda(x)$ and $u \in \operatorname{cl}\Omega \cap \operatorname{dom}\Phi$. By (15) and convexity of ρ,

$$\rho(y) - \rho(u) \le \langle g_\lambda(x) - \nabla f(x), y - u\rangle.$$

Also,

$$f(y) - f(u) = D_f(y, x) - D_f(u, x) + \langle \nabla f(x), y - u\rangle.$$

Adding these expressions and using $\langle g_\lambda(x), x - y\rangle = [D_\psi(x, y) + D_\psi(y, x)]/\lambda$ gives (4).

Under (3),

$$P_\lambda(x) = \frac{D_\psi(x, y) + D_\psi(y, x)}{\lambda} - D_f(y, x) \ge \frac{D_\psi(x, y)}{\lambda}.$$

Strict convexity of ψ on Ω gives $P_\lambda(x) = 0 \Rightarrow x = y$; the converse follows by substitution. The common point is stationary by (16). □

Corollary B.3 (Relative lower bound)**.** *Let $y \in T_\lambda(x)$. If $D_f(u, x) \ge \mu D_\psi(u, x)$ for a comparison point $u \in \operatorname{cl}\Omega \cap \operatorname{dom}\Phi \cap \operatorname{dom}\psi$, then*

$$\Phi(y) - \Phi(u) \le \langle g_\lambda(x), x - u\rangle - P_\lambda(x) - \mu D_\psi(u, x). \tag{17}$$

Proof. Substitute the lower bound in (4). □

B.3 The residual and stationarity

Proposition B.4 (Residual, descent, and stationarity)**.** *For $x \in \Omega \cap \operatorname{dom}\rho$ and $y \in T_\lambda(x)$,*

$$\mathcal{R}_\lambda(x) \ge \frac{2}{\lambda^2} D_\psi(x, y) \ge 0, \qquad \mathcal{R}_\lambda(x) = 0 \iff y = x. \tag{18}$$

If (3) holds, then

$$\Phi(y) \le \Phi(x) - \frac{\lambda}{2}\mathcal{R}_\lambda(x). \tag{19}$$

This residual vanishes exactly when $0 \in \nabla f(x) + \partial\rho(x)$.

Proof. Optimality and convexity of ρ give

$$\rho(x) - \rho(y) \ge \langle g_\lambda(x) - \nabla f(x), x - y\rangle.$$

Consequently,

$$\mathcal{R}_\lambda(x) \ge \frac{2}{\lambda}\left(\langle g_\lambda(x), x - y\rangle - \lambda^{-1} D_\psi(y, x)\right) = \frac{2}{\lambda^2} D_\psi(x, y).$$

Strict convexity and (16) give the equivalences. The identity

$$\Phi(y) - \Phi(x) = -\frac{\lambda}{2}\mathcal{R}_\lambda(x) - \lambda^{-1} D_\psi(y, x) + D_f(y, x)$$

proves (19) under the upper-bound test. □

The Bregman forward–backward envelope associated with this subproblem is

$$\mathcal{F}_\lambda(x) = f(x) + \min_{u \in \operatorname{cl}\Omega} \left\{ \rho(u) + \langle \nabla f(x), u - x \rangle + \lambda^{-1} D_\psi(u, x) \right\}.$$

For $x \in \Omega \cap \operatorname{dom}\rho$, $\mathcal{R}_\lambda(x) = \frac{2}{\lambda}(\Phi(x) - \mathcal{F}_\lambda(x))$. These identities require no differentiability assertion about $\mathcal{F}_\lambda$.

Lemma B.5 (A subgradient at the BPG point)**.** *For $y \in T_\lambda(x)$,*

$$s_\lambda(x) \coloneqq g_\lambda(x) + \nabla f(y) - \nabla f(x) \in \nabla f(y) + \partial\rho(y).$$

Proof. Add $\nabla f(y)$ to (15). □

B.4 Backtracking

Proposition B.6 (Uniform stepsize threshold)**.** *The upper-bound test holds at every admissible base point whenever $\lambda \le 1/L$. Thus geometric reductions by $\gamma_- > 1$ terminate when the base point is fixed. They also permit only finitely many stepsize reductions along any search in which the base point changes but the stepsize is never increased.*

Proof. For $y \in T_\lambda(x)$, relative smoothness gives

$$D_f(y, x) \le L D_\psi(y, x) \le \lambda^{-1} D_\psi(y, x)$$

for $\lambda \le 1/L$. The threshold is independent of x. Geometric reduction reaches it in finitely many steps, and no subsequent reduction can then be required. □

Proposition B.7 (Accepted stepsizes and backtracking count)**.** *In GA-BPGc and GA-BPGsc, initialize the first search at λ_0 and each subsequent search at $\gamma_+\lambda_{k-1}$. In GA-BPGnc, initialize at λ_0 for $k = 1$ and at*

$$\widetilde{\lambda}_k = \min\{\gamma_+\lambda_{k-1}, \lambda_{\max}\}$$

for $k \ge 2$, with $\gamma_+ \ge 1$ and $\lambda_{\max} \ge \lambda_0$. Then every accepted stepsize is at least $\underline{\lambda} = \min\{\lambda_0, 1/(\gamma_- L)\}$. For GA-BPGnc, if m_k counts failed upper-bound tests,

$$\sum_{k=1}^{N} m_k \le \frac{(N-1)\log\gamma_+ + \log(\lambda_0/\underline{\lambda})}{\log\gamma_-}. \tag{20}$$

Proof. An initial trial stepsize is no smaller than the preceding accepted value. Whenever a reduction occurs, its preceding value must exceed $1/L$; hence the reduced value exceeds $1/(\gamma_- L)$. Induction proves the lower bound, including when the base point in GA-BPGc and GA-BPGsc is recomputed after a reduction.

In GA-BPGnc, $\lambda_1 = \lambda_0\gamma_-^{-m_1}$ and $\lambda_k \le \gamma_+\lambda_{k-1}\gamma_-^{-m_k}$ for $k \ge 2$. Multiplication and the lower bound on λ_N yield (20). □

The count in (20) covers failed upper-bound tests. In GA-BPGc and GA-BPGsc, trials rejected by a κ-dependent test can require additional BPG solves; a bound on their number must be established separately.

B.5 Coefficient and mirror identities

The following results are used for GA-BPGc and GA-BPGsc, with $\phi = \psi$ in the latter.

Lemma B.8 (Coefficients and coupling)**.** *For $\omega_- \ge 0$ and $\lambda, \kappa > 0$, the convex coefficient equation has the unique positive solution*

$$\alpha = \frac{\lambda + \sqrt{\lambda^2 + 2\kappa\lambda\omega_-}}{\kappa}.$$

For $\theta_- = 1 + \mu\omega_-$ and $\mu > 0$, the strong-regime equation has a positive solution if and only if $d := \kappa - 2\mu\lambda > 0$. That solution is unique and equals

$$\alpha = \frac{\lambda(\theta_- + \mu\omega_-) + \sqrt{\lambda^2(\theta_- + \mu\omega_-)^2 + 2\lambda d\omega_-\theta_-}}{d}.$$

For fixed $\lambda, \omega_-, \theta_-$, both roots tend to zero as $\kappa \to \infty$. If $\omega = \omega_- + \alpha$ and $\omega x = \omega_- y_- + \alpha z_-$, then

$$\omega_-(x - y_-) + \alpha(x - u) = \alpha(z_- - u).$$

Proof. The respective quadratic equations are

$$\kappa\alpha^2 - 2\lambda\alpha - 2\lambda\omega_- = 0$$

and

$$d\alpha^2 - 2\lambda(\theta_- + \mu\omega_-)\alpha - 2\lambda\omega_-\theta_- = 0.$$

The quadratic formula proves the expressions and uniqueness of the positive roots, including $\omega_- = 0$. When $d \le 0$, the second polynomial is strictly negative for every $\alpha > 0$. The formulas give the limits as $\kappa \to \infty$. The coupling identity follows by expanding its left side. □

Lemma B.9 (Mirror identity)**.** *Suppose $z_-, z \in \operatorname{int}\operatorname{dom}\phi$ and $\nabla\phi(z) = \nabla\phi(z_-) - \alpha g$. For $u \in \operatorname{dom}\phi$,*

$$\alpha\langle g, z_- - u\rangle + D_\phi(u, z) - D_\phi(u, z_-) = D_\phi(z_-, z).$$

Proof. Expanding the divergences gives

$$D_\phi(u, z) - D_\phi(u, z_-) = \alpha\langle g, u - z\rangle - D_\phi(z, z_-).$$

Also,

$$\alpha\langle g, z_- - z\rangle = D_\phi(z_-, z) + D_\phi(z, z_-).$$

Add the identities. □

Lemma B.10 (Weighted mirror identity)**.** *Suppose $\theta = \theta_- + \mu\alpha$ and $z_-, x, z \in \Omega$ satisfy*

$$\theta\nabla\psi(z) = \theta_-\nabla\psi(z_-) + \mu\alpha\nabla\psi(x) - \alpha g.$$

For $u \in \operatorname{dom}\psi$,

$$\begin{aligned}\theta D_\psi(u, z) - \theta_- D_\psi(u, z_-) &= \alpha\langle g, u - z\rangle + \mu\alpha D_\psi(u, x) \\ &\quad - \mu\alpha D_\psi(z, x) - \theta_- D_\psi(z, z_-), \qquad (21)\\ \alpha\langle g, z_- - z\rangle - \mu\alpha D_\psi(z, x) - \theta_- D_\psi(z, z_-) &= \theta D_\psi(z_-, z) - \mu\alpha D_\psi(z_-, x). \qquad (22)\end{aligned}$$

Proof. Expand $\theta D_\psi(u, z) - \theta_- D_\psi(u, z_-)$. The coefficient of $\psi(u)$ is $\mu\alpha$, and the coefficient of u is $\alpha g - \mu\alpha\nabla\psi(x)$. These agree with the first two terms on the right of (21); expanding the remaining constant terms gives the other two terms. Set $u = z_-$ in that identity to obtain (22). □

Lemma B.11 (Small mirror displacement)**.** *Let ϕ be Legendre, $z_- \in \operatorname{int}\operatorname{dom}\phi$, and $p = \nabla\phi(z_-)$. If $p_\alpha = p + \alpha v_\alpha$, where $\alpha \downarrow 0$ and $\{v_\alpha\}$ is bounded, then eventually $p_\alpha \in \operatorname{int}\operatorname{dom}\phi^*$, $z_\alpha = \nabla\phi^*(p_\alpha) \to z_-$, and $D_\phi(z_-, z_\alpha) = o(\alpha)$.*

Proof. The conjugate domain is open around p, and $\nabla\phi^*$ is continuous there. Legendre duality gives $D_\phi(z_-, z_\alpha) = D_{\phi^*}(p_\alpha, p)$. Differentiability of ϕ^* at p implies $D_{\phi^*}(p_\alpha, p) = o(\|p_\alpha - p\|) = o(\alpha)$; the case $p_\alpha = p$ is immediate. □

C Auxiliary geometry and convex convergence analysis

We first discuss the choice of the auxiliary geometry ϕ. The remaining subsections provide the convergence analysis of GA-BPGc.

Table 2: Suggested pairs of BPG and auxiliary geometries and their domains.

ψ	Ω	ϕ	$\operatorname{int}\operatorname{dom}\phi^*$
$-\sum_j \log x_j$	$\mathbb{R}^n_{++}$	$\sum_j (x_j \log x_j - x_j)$	$\mathbb{R}^n$
$\frac{1}{r}\sum_j x_j^{-r}$	$\mathbb{R}^n_{++}$	$\sum_j (x_j \log x_j - x_j)$	$\mathbb{R}^n$
$-\sum_j \log(x_j(1-x_j))$	$(0,1)^n$	$\sum_j (x_j \log x_j + (1-x_j)\log(1-x_j))$	$\mathbb{R}^n$
$-\log\det X$	$\mathbb{S}^n_{++}$	$\operatorname{tr}(X\log X - X)$	$\mathbb{S}^n$
$-\log\det(X(I-X))$	$0 \prec X \prec I$	$\operatorname{tr}(X\log X + (I-X)\log(I-X))$	$\mathbb{S}^n$
$\frac{1}{q}\sum_j \lvert x_j\rvert^q$	$\mathbb{R}^n$	$\frac{1}{q}\sum_j \lvert x_j\rvert^q$	$\mathbb{R}^n$
$\frac{1}{4}\lVert x\rVert^4 + \frac{1}{2}\lVert x\rVert^2$	$\mathbb{R}^n$	$\frac{1}{4}\lVert x\rVert^4 + \frac{1}{2}\lVert x\rVert^2$	$\mathbb{R}^n$
$\sum_j h_\delta(x_j)$	$\mathbb{R}^n$	$\sum_j h_\delta(x_j)$	$\mathbb{R}^n$

C.1 Choosing the auxiliary geometry

The BPG geometry ψ is chosen to provide a relative upper bound and a tractable BPG subproblem. The auxiliary geometry ϕ is instead chosen for a computable inverse mirror map and compatibility with Ω; f need not be relatively smooth with respect to ϕ. Table 2 lists suggested pairs for GA-BPGc and GA-BPGnc. GA-BPGsc retains $\phi = \psi$ to match the geometry of relative strong convexity. These suggestions are based on domain compatibility and computational convenience, rather than a guaranteed improvement in convergence speed. We write $\mathbb{R}^n_{++}$ for the strictly positive orthant, and $\mathbb{S}^n$ and $\mathbb{S}^n_{++}$ for the sets of real symmetric and symmetric positive-definite $n \times n$ matrices, respectively. In the matrix rows, tr denotes the trace, I is the identity matrix, and $X \prec Y$ means that $Y - X$ is positive definite.

Here $r > 0$, $q > 1$, and $\delta > 0$. The hypentropy generator is $h_\delta(t) = t\operatorname{arsinh}(t/\delta) - \sqrt{t^2+\delta^2}$. For the full-space rows, $\phi(x) = \|x\|^2/2$ is an additional candidate when a simpler auxiliary update is preferred.

Domain compatibility. For the barrier–entropy pairs in Table 2, the inverse mirror maps of the auxiliary generators are defined on the full corresponding dual space and take values in the indicated domain. They are componentwise exponentials, logistic maps, or their matrix counterparts. The matrix rows use the Frobenius inner product $\langle A, B\rangle = \operatorname{tr}(A^\top B)$ and spectral matrix functions. Additional constraints encoded by ρ are not automatically preserved, so GA-BPGnc retains its check of membership in $\Omega \cap \operatorname{dom}\rho$.

Boundary comparison points and rate guarantees. With the convention $0\log 0 = 0$, the entropy generators have finite values at finite boundary points, unlike their paired barriers. This can make $D_\phi(u, z_0)$ finite for boundary comparison points in Theorem 3.5. The theorem yields a uniform accelerated rate when the accepted ratios λ_k/κ_k are bounded away from zero.

C.2 One-step estimate

For this and the remaining subsections, we consider Algorithm 2 under Assumptions 2.1, 2.2, and 3.2, with

$$y_0 = z_0 \in \Omega \cap \operatorname{dom}\rho \cap \operatorname{int}\operatorname{dom}\phi, \qquad \omega_0 = 0.$$

No comparison inequality between D_ψ and D_ϕ is imposed. Comparison points belong to $\mathcal{U}_\phi = \operatorname{cl}\Omega \cap \operatorname{dom}\Phi \cap \operatorname{dom}\phi$.

Proof of Lemma 3.3. Write $\omega_- = \omega_{k-1}$, $y_- = y_{k-1}$, and $z_- = z_{k-1}$. Apply Lemma 3.1 at the same trial (x, λ) to y_- and u:

$$\Phi(\widehat{y}) - \Phi(y_-) \le \langle g_\lambda(x), x - y_-\rangle - P_\lambda(x) - D_f(y_-, x),$$
$$\Phi(\widehat{y}) - \Phi(u) \le \langle g_\lambda(x), x - u\rangle - P_\lambda(x) - D_f(u, x).$$

Multiply these inequalities by ω_- and α, respectively. Since $\omega = \omega_- + \alpha$, the coupling identity gives

$$\begin{aligned}&\omega(\Phi(\widehat{y}) - \Phi(u)) - \omega_-(\Phi(y_-) - \Phi(u)) \\ &\quad \le \alpha\langle g_\lambda(x), z_- - u\rangle - \omega P_\lambda(x) - \omega_- D_f(y_-, x) - \alpha D_f(u, x). \end{aligned} \tag{23}$$

The mirror equation and Lemma B.9 imply

$$\alpha\langle g_\lambda(x), z_- - u\rangle + D_\phi(u, \widehat{z}) - D_\phi(u, z_-) = D_\phi(z_-, \widehat{z}).$$

Adding $D_\phi(u, \widehat{z}) - D_\phi(u, z_-)$ to both sides of (23) and using the mirror identity above gives the corresponding trial estimate. For an accepted trial, $(\omega, \alpha, x, \widehat{y}, \widehat{z}, \lambda) = (\omega_k, \alpha_k, x_k, y_k, z_k, \lambda_k)$, so the left-hand side becomes $\mathcal{E}_k^{\mathrm{C}}(u) - \mathcal{E}_{k-1}^{\mathrm{C}}(u)$. This proves (5).

Convexity makes $D_f(u, x) \ge 0$, and the acceptance test gives

$$D_\phi(z_-, \widehat{z}) - \omega P_\lambda(x) - \omega_- D_f(y_-, x) \le 0.$$

Hence every accepted iteration decreases $\mathcal{E}_k^{\mathrm{C}}(u)$. □

C.3 Telescoping and weight growth

Proof of Lemma 3.4. For every accepted iteration,

$$\sqrt{\omega_k} - \sqrt{\omega_{k-1}} = \frac{\alpha_k}{\sqrt{\omega_k} + \sqrt{\omega_{k-1}}} \ge \frac{\alpha_k}{2\sqrt{\omega_k}} = \sqrt{\frac{\lambda_k}{2\kappa_k}}.$$

Sum from $k = 1$ to N, use $\omega_0 = 0$, and square the resulting nonnegative inequality. □

Proof of Theorem 3.5. Telescoping gives $\mathcal{E}_N^{\mathrm{C}}(u) \le \mathcal{E}_0^{\mathrm{C}}(u) = D_\phi(u, z_0)$. Equivalently,

$$\omega_N(\Phi(y_N) - \Phi(u)) + D_\phi(u, z_N) \le D_\phi(u, z_0). \tag{24}$$

Since $\omega_N > 0$, dropping $D_\phi(u, z_N) \ge 0$ and using Lemma 3.4 proves (7). If $u = x^\star$ is a minimizer with finite $\phi(x^\star)$ and $\lambda_k/\kappa_k \ge c > 0$, its denominator is at least cN^2. This proves the stated conditional accelerated rate. □

The bound is meaningful for any finite comparator, not just a minimizer. If every minimizer lies outside $\operatorname{dom}\phi$, substituting a minimizer gives no finite constant. An approximation argument would then require separate control of $D_\phi(u, z_0)$ along the chosen approximations. Likewise, the bound alone does not establish that $\sum_{k=1}^N \sqrt{\lambda_k/\kappa_k}$ diverges as $N \to \infty$.

C.4 Finite termination and the stopping convention

Proposition C.1 (Finite inner search at a nonstationary iterate)**.** *Fix an outer iterate (ω_-, y_-, z_-) of GA-BPGc with $0 \notin \nabla f(y_-) + \partial\rho(y_-)$. If $\omega_- = 0$, take $y_- = z_-$ as in the initialization. Then the inner search accepts a trial or returns a stationary point after finitely many trials.*

Proof. Proposition B.6 permits only finitely many reductions of λ, even though coupling changes x. On the tail of a hypothetical infinite search, $\lambda = \bar{\lambda} > 0$ is fixed and every rejection increases κ. Thus $\kappa \to \infty$ and $\alpha \to 0$ by Lemma B.8.

If $\omega_- > 0$, then $x \to y_-$. Let $\bar{y} \in T_{\bar{\lambda}}(y_-)$. Proposition B.2 implies $\widehat{y} \to \bar{y}$, and $g_\lambda(x)$ remains bounded. The upper-bound tests on this tail and the lower bound on $P_\lambda(x)$ give

$$P_\lambda(x) \longrightarrow P_{\bar{\lambda}}(y_-) \ge \bar{\lambda}^{-1} D_\psi(y_-, \bar{y}) > 0.$$

The strict inequality uses the nonstationarity of y_-. The mirror argument tends to $\nabla\phi(z_-)$, so its domain check passes eventually and $\widehat{z} \to z_-$. Since $z_- \in \Omega$ and Ω is open, feasibility also holds eventually. Then $D_\phi(z_-, \widehat{z}) \to 0$ whereas $\omega P_\lambda(x) \to \omega_- P_{\bar{\lambda}}(y_-) > 0$. Since $D_f(y_-, x) \ge 0$, the acceptance inequality eventually holds, a contradiction.

If $\omega_- = 0$, then $x = y_- = z_-$ and $\omega = \alpha$. For the fixed $\bar{\lambda}$, $g_\lambda(x)$ and $P_\lambda(x) > 0$ are fixed. Lemma B.11 gives $D_\phi(z_-, \widehat{z}) = o(\alpha)$, whereas $\omega P_\lambda(x) = \alpha P_\lambda(x)$. The domain and feasibility checks pass eventually, and so does the acceptance inequality, again a contradiction. □

Already stationary previous iterates. Proposition C.1 does not by itself cover a stationary previous iterate, because its strict lower bound on the limiting value of $P_\lambda(x)$ then vanishes.

The stopping convention in Section 2 includes a previous-iterate check for GA-BPGc and GA-BPGsc. With λ initialized at the preceding accepted stepsize, they compute $v \in T_\lambda(y_-)$ and return y_- if $v = y_-$. By (16), this detects stationarity of y_-. If the check does not terminate, the previous iterate is nonstationary and the finite-search argument applies.

The check adds at most one BPG solve per outer iteration and leaves the trial sequence unchanged whenever the previous iterate is nonstationary.

Corollary C.2 (Finite termination with the stopping check)**.** *With this check, each outer iteration of GA-BPGc accepts a trial or returns a global minimizer after finitely many BPG solves. Without the check, Proposition C.1 and all bounds for completed iterations remain valid with their stated qualifications.*

Proof. By (16), the check detects stationarity of y_-. An interior stationary point minimizes Φ over $\operatorname{cl}\Omega$: add the convex supporting inequality for f at that point to the supporting inequality for ρ supplied by stationarity. If the check does not terminate, Proposition C.1 applies. The same argument shows that an in-trial stationary point is a global minimizer. □

C.5 Convergence under sublevel conditions

Corollary C.3 (Rate for Burg–entropy geometry)**.** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold, with $\psi(x) = -\sum_j \log x_j$ and $\phi(x) = \sum_j (x_j \log x_j - x_j)$. Assume that $\{x \in \operatorname{cl}\Omega : \Phi(x) \le r\}$ is bounded for every $r \in \mathbb{R}$, and that Φ has a minimizer in Ω. Then GA-BPGc satisfies $\sup_k \kappa_k < \infty$ and $\Phi(y_k) - \Phi^\star = \mathcal{O}(k^{-2})$. The total number of BPG solves through k completed iterations, including rejected trials and stopping checks, is $\mathcal{O}(k)$.*

Proof. Fix a minimizer $u \in \Omega$. Equation (24) gives $\Phi(y_k) \le \Phi(y_0)$ and $D_\phi(u, z_k) \le D_\phi(u, y_0)$. Hence $\{y_k\}$ is bounded, and the entropy formula places $\{z_k\}$ in a compact subset of Ω. Using the trial notation of Algorithm 2, Lemma 3.1 and the three-point identity give

$$D_\psi(u, \widehat{y}) + \lambda(\Phi(\widehat{y}) - \Phi^\star) \le D_\psi(u, x) \tag{25}$$

whenever (3) holds.

On the bounded coupling region, the directional derivative of $D_\psi(u, \cdot)$ toward z equals $\sum_j (x_j - u_j)(z_j - x_j)/x_j^2$. It is negative for sufficiently large $D_\psi(u, x)$, uniformly in z: a coordinate approaching zero contributes $-\infty$, while the other contributions are bounded above. Along coupling segments, this bounds $D_\psi(u, x)$ by the larger of $D_\psi(u, y)$ and a fixed constant. Induction using (25) therefore places all upper-bound-passing trial points $x, \widehat{y}$ in a fixed compact subset of Ω, including those subsequently rejected by the mirror criterion.

On these compact sets, the Burg formula and (23), together with the accepted Lyapunov bound, give uniformly over trials

$$\lambda \|g_\lambda(x)\|^2 = \mathcal{O}(P_\lambda(x)), \qquad \omega P_\lambda(x) = \mathcal{O}(1 + \|\alpha g_\lambda(x)\|).$$

Since $\kappa\alpha^2 = 2\lambda\omega$, it follows that $\|\alpha g_\lambda(x)\| = \mathcal{O}(\kappa^{-1/2})$ as $\kappa \to \infty$. The entropy mirror map is everywhere defined and positive, with $\widehat{z}_j = z_j e^{-\alpha g_\lambda(x)_j}$, and yields $D_\phi(z, \widehat{z}) = \mathcal{O}(\alpha^2 \|g_\lambda(x)\|^2) = \mathcal{O}(\omega P_\lambda(x)/\kappa)$. Since the preceding estimate is uniform over trials, every trial satisfying (3) with sufficiently large κ also satisfies $D_\phi(z, \widehat{z}) \le \omega P_\lambda(x)$. This implies (6) because $D_f(y, x) \ge 0$, while the entropy mirror map already satisfies both domain checks. Thus no κ-dependent rejection occurs above a fixed threshold independent of the iteration. Since the warm start never increases κ and each backtracking increase multiplies it by γ_κ, the accepted values are bounded by the larger of κ_0 and γ_κ times this threshold. Hence $\sup_k \kappa_k < \infty$.

Proposition B.7 and Theorem 3.5 now give the stated rate. Taking logarithms of the multiplicative parameter updates, using the stepsize lower bound and the bound on κ_k, shows that the total number of rejections is $\mathcal{O}(k)$. Including successful trials and stopping checks proves the BPG-solve bound. □

Algorithm 3: GA-BPGsc

Input: Known $\mu > 0$; $y_0 \in \Omega \cap \operatorname{dom} \rho$; $\lambda_0, \kappa_0 > 0$; $\gamma_+ \geq 1$; $\gamma_-, \gamma_\kappa > 1$.

$z_0 \leftarrow y_0, \omega_0 \leftarrow 0, \theta_0 \leftarrow 1$

for $k = 1, 2, \ldots$ **do**

  $(\widehat{\lambda}, \widehat{\kappa}) \leftarrow (\gamma_+ \lambda_{k-1}, \kappa_{k-1}/\gamma_+)$

  $(\alpha_k, \omega_k, \theta_k, x_k, y_k, z_k, g_k, \lambda_k, \kappa_k) \leftarrow \texttt{Backtracking}\,(\omega_{k-1}, \theta_{k-1}, y_{k-1}, z_{k-1}, \widehat{\lambda}, \widehat{\kappa})$

Procedure `Backtracking` $(\omega_-, \theta_-, y, z, \lambda, \kappa)$

  if $\kappa \leq 2\mu\lambda$ **then** ▷ `Criterion`

    return `Backtracking` $(\omega_-, \theta_-, y, z, \lambda, \gamma_\kappa \kappa)$

  $d \leftarrow \kappa - 2\mu\lambda$, $\alpha \leftarrow [\lambda(\theta_- + \mu\omega_-) + \sqrt{\lambda^2(\theta_- + \mu\omega_-)^2 + 2\lambda d \omega_- \theta_-}]/d$ ▷ `Coupling`

  $\omega \leftarrow \omega_- + \alpha$, $\theta \leftarrow \theta_- + \mu\alpha$, $x \leftarrow (\omega_- y + \alpha z)/\omega$

  $\widehat{y} \in T_\lambda(x)$ ▷ `BPG update`

  if $D_f(\widehat{y}, x) > D_\psi(\widehat{y}, x)/\lambda$ **then**

    return `Backtracking` $(\omega_-, \theta_-, y, z, \lambda/\gamma_-, \kappa)$

  $g_\lambda(x) \leftarrow [\nabla\psi(x) - \nabla\psi(\widehat{y})]/\lambda$

  $P_\lambda(x) \leftarrow [D_\psi(x, \widehat{y}) + D_\psi(\widehat{y}, x)]/\lambda - D_f(\widehat{y}, x)$

  $p \leftarrow [\theta_- \nabla\psi(z) + \mu\alpha\nabla\psi(x) - \alpha g_\lambda(x)]/\theta$

  if $p \notin \operatorname{int} \operatorname{dom} \psi^*$ **then**

    return `Backtracking` $(\omega_-, \theta_-, y, z, \lambda, \gamma_\kappa \kappa)$

  $\widehat{z} \leftarrow \nabla\psi^*(p)$ ▷ `Mirror update`

  if $\theta D_\psi(z, \widehat{z}) - \mu(\omega_- D_\psi(y, x) + \alpha D_\psi(z, x)) > \omega P_\lambda(x)$ **then** ▷ `Criterion`

    return `Backtracking` $(\omega_-, \theta_-, y, z, \lambda, \gamma_\kappa \kappa)$

  return $(\alpha, \omega, \theta, x, \widehat{y}, \widehat{z}, g_\lambda(x), \lambda, \kappa)$

C.6 Euclidean specialization

Corollary C.4 (Euclidean rate with fixed κ)**.** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold, $\psi = \phi = \|\cdot\|^2/2$, and $\Omega = \mathbb{R}^n$. Run GA-BPGc with $\gamma_+ = 1$ and $\kappa_0 \geq 2$, and set $\underline{\lambda} = \min\{\lambda_0, 1/(\gamma_- L)\}$. Then $\lambda_k \geq \underline{\lambda}$ and $\kappa_k = \kappa_0$ at every completed iteration. For every $N \geq 1$ and $u \in \operatorname{dom} \Phi$,*

$$\Phi(y_N) - \Phi(u) \leq \frac{\kappa_0 \|u - z_0\|^2}{\underline{\lambda} N^2}.$$

The number of BPG solves through N completed iterations is $\mathcal{O}(N)$.

Proof. Consider a trial with $\kappa = \kappa_0$ satisfying (3), and write $g = g_\lambda(x)$. In Euclidean geometry,

$$P_\lambda(x) \geq \frac{\lambda}{2}\|g\|^2, \qquad D_\phi(z_-, \widehat{z}) = \frac{\alpha^2}{2}\|g\|^2 = \frac{\lambda\omega}{\kappa_0}\|g\|^2 \leq \omega P_\lambda(x).$$

Since $D_f(y_-, x) \geq 0$, the acceptance test (6) holds. The domain checks are vacuous. Thus no κ increase occurs, and $\gamma_+ = 1$ gives $\kappa_k = \kappa_0$. Proposition B.7 gives the stepsize lower bound, and Theorem 3.5 gives the stated rate. Only finitely many stepsize reductions occur because stepsizes are never increased and every trial with $\lambda \leq 1/L$ passes the upper-bound test. The stopping convention adds at most one BPG solve per outer iteration, proving the cost bound. □

D Convergence analysis for relatively strongly convex objectives

We impose the relatively strongly convex condition of Assumption 3.6, with known $\mu > 0$. Comparison points belong to $\mathcal{U}_\psi = \operatorname{cl}\Omega \cap \operatorname{dom}\Phi \cap \operatorname{dom}\psi$. The BPG and mirror updates both use ψ. The stopping convention in Section 2 applies; its tests are omitted from the pseudocode.

D.1 Algorithm

The acceptance test is given by (9). The conjugate-domain check is performed before the inverse mirror map; that map returns a point in Ω. A rejected trial retains the previous accepted points and

weights, as well as any changes to the trial λ and κ. Each recursive call retries with these updated trial parameters. Induction gives $\theta_k = 1 + \mu\omega_k$.

D.2 Lyapunov decrease and the product bound

We use the Lyapunov function $\mathcal{E}_k^{\mathrm{S}}(u)$ defined in (8).

Proof of Lemma 3.7. Consider a trial satisfying the coupling and mirror-update equations of GA-BPGsc, with $\widehat{y} \in T_\lambda(x)$ and a valid mirror candidate $\widehat{z} \in \Omega$, and set $\omega = \omega_{k-1} + \alpha$ and $\theta = \theta_{k-1} + \mu\alpha$. Apply (17) with comparison points y_{k-1} and u, multiply by ω_{k-1} and α, respectively, and add. The coupling identity gives

$$\begin{aligned}&\omega(\Phi(\widehat{y}) - \Phi(u)) - \omega_{k-1}(\Phi(y_{k-1}) - \Phi(u))\\ &\quad\le \alpha\langle g_\lambda(x), z_{k-1} - u\rangle - \omega P_\lambda(x) - \mu\omega_{k-1} D_\psi(y_{k-1}, x) - \mu\alpha D_\psi(u, x).\end{aligned}$$

The weighted mirror identities (21) and (22) imply

$$\begin{aligned}&\alpha\langle g_\lambda(x), z_{k-1} - u\rangle + \theta D_\psi(u, \widehat{z}) - \theta_{k-1} D_\psi(u, z_{k-1})\\ &\quad= \mu\alpha D_\psi(u, x) + \theta D_\psi(z_{k-1}, \widehat{z}) - \mu\alpha D_\psi(z_{k-1}, x).\end{aligned}$$

Adding the change in weighted Bregman divergence to the objective estimate cancels the terms containing $D_\psi(u, x)$. For an accepted trial, $(\omega, \theta, \alpha, x, \widehat{y}, \widehat{z}, \lambda) = (\omega_k, \theta_k, \alpha_k, x_k, y_k, z_k, \lambda_k)$, so the left-hand side becomes $\mathcal{E}_k^{\mathrm{S}}(u) - \mathcal{E}_{k-1}^{\mathrm{S}}(u)$, which yields (10). □

Proof of (11). The coefficient condition gives $\kappa_k > 2\mu\lambda_k$, hence $\sqrt{2\mu\lambda_k/\kappa_k} \in (0, 1)$. Since $\theta_k = 1 + \mu\omega_k \ge \mu\omega_k$, the coefficient equation implies

$$\alpha_k = \sqrt{\frac{2\lambda_k\omega_k\theta_k}{\kappa_k}} \ge \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\,\omega_k.$$

Therefore

$$\omega_{k-1} = \omega_k - \alpha_k \le \left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right)\omega_k.$$

Iterating this inequality from $k = 2$ to N proves (11). For $N = 1$, the assertion holds with an empty product. □

Proof of Theorem 3.8. Fix $u \in \mathcal{U}_\psi$. Lemma 3.7 and the acceptance test give

$$\mathcal{E}_k^{\mathrm{S}}(u) \le \mathcal{E}_{k-1}^{\mathrm{S}}(u)$$

at every completed iteration. Since $\omega_0 = 0$ and $\theta_0 = 1$, telescoping yields

$$\mathcal{E}_N^{\mathrm{S}}(u) \le \mathcal{E}_0^{\mathrm{S}}(u) = D_\psi(u, z_0).$$

Equivalently,

$$\omega_N(\Phi(y_N) - \Phi(u)) + \theta_N D_\psi(u, z_N) \le D_\psi(u, z_0). \tag{26}$$

Dropping $\theta_N D_\psi(u, z_N) \ge 0$ and using (11), we obtain

$$\Phi(y_N) - \Phi(u) \le \frac{D_\psi(u, z_0)}{\omega_N} \le \frac{D_\psi(u, z_0)}{\omega_1} \prod_{k=2}^{N} \left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right),$$

which proves (12). Taking $u = x^\star$ gives the optimality-gap bound. If $\sqrt{2\mu\lambda_k/\kappa_k} \ge q_{\min} > 0$, then

$$\prod_{k=2}^{N} \left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right) \le (1 - q_{\min})^{N-1},$$

so the objective gap converges linearly. □

Corollary D.1 (Distance statements and boundary points)**.** *For a minimizer $x^\star \in \mathcal{U}_\psi$,*

$$D_\psi(x^\star, z_N) \le \frac{D_\psi(x^\star, y_0)}{\theta_N}. \tag{27}$$

If, in addition, $x^\star \in \Omega$, then

$$D_\psi(y_N, x^\star) \le \frac{D_\psi(x^\star, y_0)}{\mu\omega_1} \prod_{k=2}^{N} \left(1 - \sqrt{\frac{2\mu\lambda_k}{\kappa_k}}\right). \tag{28}$$

Proof. In (26), the objective gap is nonnegative for a minimizer. Dropping it proves (27).

For an interior minimizer, first-order optimality gives $-\nabla f(x^\star) \in \partial\rho(x^\star)$. Add the supporting inequality for ρ to relative strong convexity at $x^\star$ to obtain

$$\Phi(y) - \Phi^\star \ge \mu D_\psi(y, x^\star).$$

The product bound gives (28). □

D.3 Finite termination

Proposition D.2 (Finite strong-regime search)**.** *At a nonstationary previous iterate, the inner search of Algorithm 3 terminates finitely. With its previous-iterate stopping check, every outer iteration therefore accepts a trial or returns a global minimizer after finitely many BPG solves.*

Proof. The stopping check detects a stationary $y_- = y_{k-1}$ by (16). Such a point is a global minimizer by the relative lower bound and convexity of ρ. Otherwise y_- is nonstationary.

There can be only finitely many reductions of λ. On an infinite remaining tail, $\lambda = \bar\lambda$ is fixed and $\kappa \to \infty$. The coefficient condition eventually holds, and $\alpha \to 0$, $\omega \to \omega_-$, $\theta \to \theta_-$.

For $\omega_- > 0$, let $\bar y \in T_{\bar\lambda}(y_-)$. Then $x \to y_-$, and the continuity and positivity arguments in Appendix C.4 give $P_\lambda(x) \to P_{\bar\lambda}(y_-) > 0$. The mirror argument converges to $\nabla\psi(z_-)$, so it is eventually in the conjugate domain and $\widehat{z} \to z_-$. Since the subtracted divergences in (9) are nonnegative,

$$\begin{aligned}\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right)\\ \le \theta D_\psi(z_-, \widehat{z}) \to 0,\end{aligned}$$

whereas $\omega P_\lambda(x) \to \omega_- P_{\bar\lambda}(y_-) > 0$. Acceptance follows eventually.

For $\omega_- = 0$, initialization gives $x = y_- = z_-$ and $\theta_- = 1$. The mirror argument is $\nabla\psi(z_-) - (\alpha/\theta)g_\lambda(x)$. Lemma B.11 with $\phi = \psi$ yields

$$\begin{aligned}\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right)\\ = \theta D_\psi(z_-, \widehat{z}) = o(\alpha),\end{aligned}$$

while $\omega P_\lambda(x) = \alpha P_\lambda(x) > 0$. The domain and acceptance tests again hold eventually. These contradictions prove finite termination. □

D.4 Convergence under sublevel conditions

Corollary D.3 (Linear rate for Burg geometry)**.** *Suppose that Assumptions 2.1, 2.2 and 3.6 hold, with $\psi(x) = -\sum_j \log x_j$ on $\Omega = \mathbb{R}^n_{++}$. Then Φ has a unique minimizer in Ω, and GA-BPGsc satisfies $\sup_k \kappa_k < \infty$ and $\Phi(y_k) - \Phi^\star = \mathcal{O}(q^k)$ for some $q \in (0,1)$. The total number of BPG solves through k completed iterations, including rejected trials and stopping checks, is $\mathcal{O}(k)$.*

Proof. Assumption 3.6 makes $f - \mu\psi$ convex on Ω and excludes finite objective values on $\partial\Omega$. For any $r \in \mathbb{R}$ and $u, v \in \Omega$ with $\Phi(u), \Phi(v) \le r$, the midpoint inequality gives

$$\sum_j \log \frac{(u_j + v_j)^2}{4u_j v_j} \le \frac{2(r - \Phi^\star)}{\mu}.$$

Fixing u in any nonempty sublevel, each summand is nonnegative and diverges as $v_j \downarrow 0$ or $v_j \to \infty$. Thus every objective sublevel is compact and contained in Ω. Lower semicontinuity and strict convexity therefore yield a unique minimizer $u \in \Omega$.

Equation (26) gives $\Phi(y_k) \le \Phi(y_0)$ and $D_\psi(u, z_k) \le D_\psi(u, y_0)$. Since the sublevels of $D_\psi(u, \cdot)$ are compact in Ω, both $\{y_k\}$ and $\{z_k\}$ lie in fixed compact subsets of Ω. Using the trial notation of Algorithm 3, coupling and (25) therefore place all upper-bound-passing trial points $x, z, \widehat{y}$ in a fixed compact subset of Ω, including those subsequently rejected by the mirror criterion.

For a nonterminal trial satisfying (3), relative strong convexity gives $\mu\lambda \le 1$. Since $\theta = 1 + \mu\omega$ and $\kappa\alpha^2 = 2\lambda\omega\theta$, we have $\mu\alpha/\theta \le \sqrt{2\mu\lambda/\kappa} \le \sqrt{2/\kappa}$. The stepsize-reduction argument in Proposition B.7 also bounds all trial stepsizes away from zero, so $g_\lambda(x)$ is uniformly bounded. The mirror argument is

$$p = \nabla\psi(z) - \frac{\alpha}{\theta}\left(g_\lambda(x) + \mu(\nabla\psi(z) - \nabla\psi(x))\right).$$

Hence $p - \nabla\psi(z) = \mathcal{O}(\kappa^{-1/2})$ uniformly over trials. Because $\nabla\psi(z)$ lies in a fixed compact subset of $\operatorname{int}\operatorname{dom}\psi^*$, the conjugate-domain check passes for all sufficiently large κ, and $\widehat{z}$ remains in a fixed compact subset of Ω.

On these compact sets, the Burg formula gives $\lambda\|g_\lambda(x)\|^2 = \mathcal{O}(P_\lambda(x))$ and $\|\nabla\psi(z) - \nabla\psi(x)\|^2 = \mathcal{O}(D_\psi(z, x))$, uniformly over trials. The logarithmic remainder estimate for the mirror update, together with the coefficient equation, consequently yields

$$\begin{aligned}\theta D_\psi(z, \widehat{z}) &= \mathcal{O}\left(\frac{\alpha^2}{\theta}\|g_\lambda(x)\|^2 + \frac{\mu^2\alpha^2}{\theta}D_\psi(z, x)\right)\\ &= \mathcal{O}\left(\frac{\omega P_\lambda(x)}{\kappa}\right) + \mathcal{O}\left(\frac{\mu\alpha}{\theta}\right)\mu\alpha D_\psi(z, x).\end{aligned}$$

Since $1/\kappa$ and $\mu\alpha/\theta$ both tend to zero uniformly, every upper-bound-passing trial with sufficiently large κ satisfies $\theta D_\psi(z, \widehat{z}) \le \omega P_\lambda(x) + \mu\alpha D_\psi(z, x)$. This implies (9), because the remaining term $-\mu\omega_- D_\psi(y, x)$ is nonpositive.

The coefficient check also passes above a uniform threshold: accepted nonterminal stepsizes satisfy $\mu\lambda_k \le 1$, so every trial has $\lambda \le \gamma_+ \max\{\lambda_0, 1/\mu\}$. Thus no κ-dependent rejection occurs above a fixed threshold independent of the iteration. Since the warm start never increases κ and each backtracking increase multiplies it by γ_κ, the accepted values are bounded by the larger of κ_0 and γ_κ times this threshold. Hence $\sup_k \kappa_k < \infty$.

Proposition B.7 and Theorem 3.8 now give the stated linear rate. Taking logarithms of the multiplicative parameter updates, using the stepsize lower bound and the bound on κ_k, shows that the total number of rejections is $\mathcal{O}(k)$. Each rejection requires at most one BPG solve; adding the accepted trials and previous-iterate stopping checks proves the BPG-solve bound. □

D.5 Euclidean specialization

Corollary D.4 (Euclidean linear rate with fixed κ)**.** *Suppose that Assumptions 2.1, 2.2 and 3.6 hold, $\psi = \|\cdot\|^2/2$, and $\Omega = \mathbb{R}^n$. Run GA-BPGsc with $\gamma_+ = 1$, $\kappa_0 \ge 8$, and $\kappa_0 > 2\mu\lambda_0$. Set*

$$\underline{\lambda} = \min\{\lambda_0, 1/(\gamma_- L)\}, \qquad q_{\min} = \sqrt{\frac{2\mu\underline{\lambda}}{\kappa_0}}.$$

Then $\lambda_k \ge \underline{\lambda}$ and $\kappa_k = \kappa_0$ at every completed iteration. For any minimizer $x^\star$ and $N \ge 1$,

$$\Phi(y_N) - \Phi^\star \le \frac{\|x^\star - y_0\|^2}{2\omega_1}(1 - q_{\min})^{N-1}.$$

Proof. The stepsize lower bound has already been proved. For a trial with $\omega_- > 0$, put

$$d = z_- - x, \qquad r = \alpha/\omega_-.$$

The coefficient condition gives $\sqrt{2\mu\lambda/\kappa} \in (0,1)$. The coupling and mirror equations give $x - y_- = rd$ and $\theta(z_- - \widehat{z}) = \alpha(g_\lambda(x) + \mu d)$. Expand the left-hand side of (9) and use $\alpha^2 = 2\lambda\omega\theta/\kappa$:

$$\begin{aligned}&\frac{\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right)}{\omega} \\ &\quad = \frac{\lambda}{\kappa}\|g_\lambda(x)\|^2 + \frac{2\mu\lambda}{\kappa}\langle g_\lambda(x), d\rangle - \frac{\mu}{2}\left(r - \frac{2\mu\lambda}{\kappa}\right)\|d\|^2.\end{aligned} \tag{29}$$

Since $\theta \geq \mu\omega$, $\alpha/\omega \geq \sqrt{2\mu\lambda/\kappa}$. As $\alpha/\omega = r/(1+r)$, this implies

$$r \geq \frac{\sqrt{2\mu\lambda/\kappa}}{1 - \sqrt{2\mu\lambda/\kappa}}, \qquad \frac{2\mu\lambda}{\kappa r} \leq \sqrt{\frac{2\mu\lambda}{\kappa}}\left(1 - \sqrt{\frac{2\mu\lambda}{\kappa}}\right) \leq \frac{1}{4}.$$

Completing the square in d in (29) gives

$$\begin{aligned}&\frac{\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right)}{\omega} \\ &\quad \leq \frac{\lambda}{\kappa}\frac{r}{r - 2\mu\lambda/\kappa}\|g_\lambda(x)\|^2 \leq \frac{4\lambda}{3\kappa}\|g_\lambda(x)\|^2.\end{aligned}$$

The upper-bound test gives $P_\lambda(x) \geq \lambda\|g_\lambda(x)\|^2/2$. Thus

$$\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right) \leq \omega P_\lambda(x)$$

whenever $\kappa \geq 8$.

For $\omega_- = 0$, $x = y_- = z_-$ and

$$\begin{aligned}&\theta D_\psi(z_-, \widehat{z}) - \mu\left(\omega_- D_\psi(y_-, x) + \alpha D_\psi(z_-, x)\right) \\ &\quad = \frac{\alpha^2}{2\theta}\|g_\lambda(x)\|^2 = \frac{\lambda\omega}{\kappa}\|g_\lambda(x)\|^2,\end{aligned}$$

so $\kappa \geq 2$ suffices. The mirror-domain tests are vacuous in Euclidean geometry.

Since $\gamma_+ = 1$, every trial satisfies $\lambda \leq \lambda_0$. The condition $\kappa_0 > 2\mu\lambda_0$ prevents coefficient rejection, and $\kappa_0 \geq 8$ ensures acceptance after the upper-bound test. Hence no κ increase occurs and $\kappa_k = \kappa_0$. Proposition B.7 gives $\lambda_k \geq \underline{\lambda}$. Substitution in Theorem 3.8 proves the rate. □

For example, fixed $\lambda = 1/L$ and $\kappa = 8$ give contraction factor $1 - \frac{1}{2}\sqrt{\mu/L}$. As in the convex Euclidean case, the total number of parameter rejections is bounded under the prescribed warm starts; including the stopping check still gives $\mathcal{O}(N)$ BPG solves through N outer iterations.

E Convergence analysis for nonconvex objectives

We impose the common assumptions, including $\Phi^\star > -\infty$. The analysis below uses no relative lower bound or DC decomposition and does not assume $\Omega = \mathbb{R}^n$. Unlike in GA-BPGc and GA-BPGsc, the variable z_k stores the trial point for the next iteration. The fixed auxiliary geometry ϕ is used only to construct that point, not in the stationarity bound.

E.1 Algorithm and well-definedness

Backtracking holds x_k fixed. Every early return in the criterion completes the outer iteration with $z_k = y_k$. Thus neither an invalid gradient nor an invalid inverse mirror map is evaluated. Each weight β_{k+1} is chosen from the history available before backtracking and held fixed throughout its trials.

Proposition E.1 (Well-defined outer iterations)**.** *Every nonterminal outer iteration is completed after finitely many BPG trials. Its variables satisfy $x_k, \widehat{y}_k, y_k, z_k \in \Omega \cap \operatorname{dom}\rho$, and $\mathcal{R}_{\lambda_k}(x_k) > 0$.*

Proof. Initially $z_0 = y_0 \in \Omega \cap \operatorname{dom}\rho$. Inductively, $x_k = z_{k-1}$ is feasible and interior. Solvability places every BPG solution in the same set, and Proposition B.6 gives finite backtracking. If the stopping test fails, Proposition B.4 gives $\mathcal{R}_{\lambda_k}(x_k) > 0$. The point y_k is either $\widehat{y}_k$ or y_{k-1}. The default $z_k = y_k$ is replaced only by a candidate that passes the explicit feasibility test, closing the induction. □

Algorithm 4: GA-BPGnc

Input: A Legendre function ϕ; $y_0 \in \Omega \cap \operatorname{dom} \rho$; $\lambda_{\max} \geq \lambda_0 > 0$; $\gamma_+ \geq 1$, $\gamma_- > 1$; $\sigma \in (0,1)$; a rule for choosing weights $\beta_j \geq 0$.

$z_0 \leftarrow y_0$
for $k = 1, 2, \ldots$ **do**
  if $k = 1$ **then**
    $\lambda \leftarrow \lambda_0$
  else
    $\lambda \leftarrow \min\{\gamma_+ \lambda_{k-1}, \lambda_{\max}\}$
  Choose $\beta_{k+1} \geq 0$ by the prescribed rule
  $(x_k, \widehat{y}_k, y_k, z_k, \lambda_k, \delta_k) \leftarrow$ `Backtracking` $(y_{k-1}, z_{k-1}, \lambda, \beta_{k+1})$
Procedure `Backtracking` (y, z, λ, β)
  $x \leftarrow z$ ▷ `Coupling`
  $\widehat{y} \in T_\lambda(x)$ ▷ `BPG update`
  if $D_f(\widehat{y}, x) > D_\psi(\widehat{y}, x)/\lambda$ **then**
    return `Backtracking` $(y, z, \lambda/\gamma_-, \beta)$
  $\delta \leftarrow \Phi(y) - \Phi(\widehat{y})$
  if $\delta > 0$ **then**
    $y_+ \leftarrow \widehat{y}$
  else
    $y_+ \leftarrow y$
  $z_+ \leftarrow y_+$
  if $\delta < \frac{\sigma\lambda}{2} \mathcal{R}_\lambda(x)$ **then**
    return $(x, \widehat{y}, y_+, z_+, \lambda, \delta)$
  if $\{y_+, y\} \not\subset \operatorname{int} \operatorname{dom} \phi$ **then**
    return $(x, \widehat{y}, y_+, z_+, \lambda, \delta)$
  $p \leftarrow \nabla\phi(y_+) + \beta[\nabla\phi(y_+) - \nabla\phi(y)]$
  if $p \notin \operatorname{int} \operatorname{dom} \phi^*$ **then**
    return $(x, \widehat{y}, y_+, z_+, \lambda, \delta)$
  $\widehat{z} \leftarrow \nabla\phi^*(p)$ ▷ `Mirror update`
  if $\widehat{z} \in \Omega \cap \operatorname{dom} \rho$ **then** ▷ `Criterion`
    $z_+ \leftarrow \widehat{z}$
  return $(x, \widehat{y}, y_+, z_+, \lambda, \delta)$

E.2 Descent and reset

Proof of Lemma 3.9. The acceptance rule immediately gives (13). For $x_k = y_{k-1}$, inequality (19) gives $\delta_k \geq \frac{\lambda_k}{2} \mathcal{R}_{\lambda_k}(x_k) > 0$, hence $\tau_k \geq 1$. If $\tau_k < \sigma$, then $\delta_k < \frac{\sigma\lambda_k}{2} \mathcal{R}_{\lambda_k}(x_k)$, so the algorithm keeps $z_k = y_k$. Consequently $x_{k+1} = y_k$, and the preceding argument applies to the next iteration. An invalid extrapolation also gives a safe next iteration, although this additional case is not needed for the bound. □

E.3 Stationarity bound in terms of objective decreases and stepsizes

Proof of the first inequality in (14). The first iteration is safe because $x_1 = z_0 = y_0$. If it is nonterminal, Lemma 3.9 gives $\tau_1 \geq 1$, so $\mathcal{A}_N$ is nonempty. Telescope (13):

$$\Phi(y_0) - \Phi(y_N) = \frac{1}{2} \sum_{k \in \mathcal{A}_N} \tau_k \lambda_k \mathcal{R}_{\lambda_k}(x_k).$$

All weights in this sum are positive. Since $y_N \in \operatorname{cl} \Omega$, its left side is at most $\Phi(y_0) - \Phi^\star$. Taking the minimum residual outside the sum and dividing by the sum of weights proves the first bound in (14). □

E.4 WORST-CASE BOUND

Lemma E.2 (Two-iteration inequality)**.** *If both iterations k and $k+1$ are nonterminal, then*

$$[\tau_k]_+ + [\tau_{k+1}]_+ \geq \sigma.$$

Proof. If $\tau_k \geq \sigma$, its first term suffices. Otherwise, the next iteration is safe and $\tau_{k+1} \geq 1 > \sigma$ by Lemma 3.9. □

Proof of the second inequality in (14). Summing Lemma E.2 over $k = 1, \ldots, N-1$ gives

$$2\sum_{k=1}^{N}[\tau_k]_+ \geq \sigma(N-1).$$

The stepsize lower bound therefore implies

$$\sum_{k\in\mathcal{A}_N} \tau_k\lambda_k \geq \underline{\lambda}\sum_{k=1}^{N}[\tau_k]_+ \geq \tfrac{1}{2}\sigma\underline{\lambda}(N-1).$$

Substitute this lower bound into the first inequality in (14). A residual at most ε^2 thus requires at most

$$1 + \left\lceil \frac{4(\Phi(y_0) - \Phi^\star)}{\sigma\underline{\lambda}\varepsilon^2} \right\rceil$$

outer iterations, unless stationarity is returned earlier. □

Corollary E.3 (A sharper counting bound)**.** *With $c_\sigma = \min\{\sigma, 1/2\}$, after any $N \geq 1$ nonterminal iterations,*

$$\sum_{k=1}^{N}[\tau_k]_+ \geq c_\sigma N, \qquad \min_{k\in\mathcal{A}_N} \mathcal{R}_{\lambda_k}(x_k) \leq \frac{2(\Phi(y_0) - \Phi^\star)}{c_\sigma\underline{\lambda}N}.$$

Proof. Call an iteration bad when $\tau_k < \sigma$. The first iteration is not bad, and each bad iteration other than the last is followed by a safe iteration with $\tau \geq 1$. Pair each such bad iteration with its successor. If the last iteration is bad, pair it with the first iteration instead. No pairs overlap, since bad iterations cannot be consecutive and the first iteration is safe.

If there are b bad iterations, each of the b pairs contributes at least one to the sum, and each of the remaining $N - 2b$ iterations contributes at least σ. Therefore the sum is at least

$$b + \sigma(N - 2b) \geq \min\{\sigma, 1/2\}N.$$

Combine this with the stepsize lower bound and the first inequality in (14). □

E.5 NUMBER OF BPG SOLVES

Corollary E.4 (Amortized backtracking cost)**.** *Through N nonterminal outer iterations,*

$$N_{\mathrm{BPG}}(N) \leq N + \frac{(N-1)\log\gamma_+ + \log(\lambda_0/\underline{\lambda})}{\log\gamma_-}.$$

Consequently, the BPG-solve complexity for $\mathcal{R}_{\lambda_k}(x_k) \leq \varepsilon^2$ is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Each outer iteration performs one successful upper-bound trial and m_k failed trials, so

$$N_{\mathrm{BPG}}(N) = N + \sum_{k=1}^{N} m_k.$$

Apply (20) and the outer-iteration bound. A terminal iteration satisfies the same backtracking estimate with its actual number of trials. Constructing the next mirror candidate requires no additional BPG subproblem solve. □

E.6 Relation to norm-based stationarity

We use $\operatorname{dist}(x, S) := \inf_{u \in S} \|x - u\|$ for the Euclidean distance from x to a set S.

Proposition E.5 (Subgradient comparison)**.** *Let*

$$s_k = g_{\lambda_k}(x_k) + \nabla f(\widehat{y}_k) - \nabla f(x_k).$$

Then $s_k \in \nabla f(\widehat{y}_k) + \partial\rho(\widehat{y}_k)$. If a uniform constant $\Gamma > 0$ satisfies $\|s_k\|^2 \le \Gamma \mathcal{R}_{\lambda_k}(x_k)$ for $k \in \mathcal{A}_N$, then

$$\min_{k \in \mathcal{A}_N} \operatorname{dist}^2(0, \nabla f(\widehat{y}_k) + \partial\rho(\widehat{y}_k)) \le \frac{4\Gamma(\Phi(y_0) - \Phi^\star)}{\sigma \underline{\lambda}(N-1)}, \qquad N \ge 2.$$

Proof. Membership follows from Lemma B.5. The squared distance is at most $\|s_k\|^2$. Apply the assumed uniform comparison and the second inequality in (14). □

The uniform comparison with Γ is an additional assumption beyond upper relative smoothness.

Corollary E.6 (Euclidean residual)**.** *If $\psi = \|\cdot\|^2/2$ and $\Omega = \mathbb{R}^n$, then, for every $x \in \operatorname{dom}\rho$ and $y \in T_\lambda(x)$,*

$$\left\| \frac{x - y}{\lambda} \right\|^2 \le \mathcal{R}_\lambda(x).$$

If also $\rho = 0$, then $\mathcal{R}_\lambda(x) = \|\nabla f(x)\|^2$. Thus the nonconvex main-text bounds recover the corresponding bounds on the squared gradient norm in the smooth unconstrained case.

Proof. The first assertion follows from (18) with $D_\psi(x, y) = \|x - y\|^2/2$. When $\rho = 0$, $T_\lambda(x) = \{x - \lambda\nabla f(x)\}$, and direct substitution gives $\mathcal{R}_\lambda(x) = \|\nabla f(x)\|^2$. □

The stationarity bounds concern interior evaluation points. Convergence of the full sequence and stationarity at boundary accumulation points require separate analysis.

F Additional numerical experiments and more discussions

F.1 Additional experiments

Convex D-optimal design. Following Hanzely et al. (2021), we solve $\min_{x \in \Delta_n} \Phi(x) := -\log\det(H \operatorname{Diag}(x) H^\top)$, where $\operatorname{Diag}(x)$ is the diagonal matrix with diagonal x and $\Delta_n = \{x \in \mathbb{R}^n : x \ge 0,\ \mathbf{1}^\top x = 1\}$. The columns of $H \in \mathbb{R}^{m \times n}$ are the raw LIBSVM `abalone` feature vectors ($m = 8$, $n = 4{,}177$). All methods use the Burg geometry $\psi(x) = -\sum_i \log x_i$ and start at $x_0 = \mathbf{1}/n$; GA-BPGc uses the auxiliary negative entropy $\phi(x) = \sum_i (x_i \log x_i - x_i)$. Figure 4 compares GA-BPGc, BPG-LS, and ABPG-g over five timing runs, each capped at 5,000 iterations or 1,800 seconds, showing medians and interquartile ranges. The deterministic iteration traces coincide across repetitions. Separate long runs of BPG, BPG-LS, and ABPG-g provide x_{ref} and the certified lower bound $\Phi_{\mathrm{LB}} = \Phi(x_{\mathrm{ref}}) - \langle \nabla\Phi(x_{\mathrm{ref}}), x_{\mathrm{ref}} \rangle + \min_j [\nabla\Phi(x_{\mathrm{ref}})]_j$. We use $\Phi_{\mathrm{ref}} = \Phi_{\mathrm{LB}}$. The dotted curve uses the right-hand side of (7) with $u = x_{\mathrm{ref}}$, adding $\Phi(x_{\mathrm{ref}}) - \Phi_{\mathrm{LB}}$ before applying the same normalization.

Nonconvex phase retrieval. We consider the normalized ℓ_1-regularized phase-retrieval problem $\min_{x \in \mathbb{R}^d} \Psi(x) := \frac{1}{4m} \sum_{r=1}^m \left(\langle a_r, x \rangle^2 - b_r \right)^2 + \frac{1}{m} \|x\|_1$. We generate five independent instances (five seeds) with $d = 1000$, $m = 6000$, independent $a_r \sim \mathcal{N}(0, I)$, and noiseless measurements $b_r = \langle a_r, x^\star \rangle^2$ of a 5%-sparse Gaussian signal. All algorithms share each instance's spectral initializer x_0 (Candès et al., 2015). We compare GA-BPGnc with BPG, BPGe, CoCaIn-BPG, and BPDCAe (Takahashi et al., 2022). All algorithms use $\psi(x) = \phi(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$. BPG and BPGe use $\lambda = 0.1/L$ with a Gaussian-based estimate L; BPDCAe uses the deterministic bound (Takahashi et al., 2022). GA-BPGnc and BPDCAe use $\beta_k = (t_{k-1} - 1)/t_k$ and $t_k = (1 + (1 + 4t_{k-1}^2)^{1/2})/2$. GA-BPGnc and CoCaIn-BPG use their respective backtracking rules. Runs are capped at 2,000 iterations; only GA-BPGnc stops early upon meeting its stationarity criterion. Figure 5 plots normalized empirical objective gaps relative to Ψ_{ref}. Across all five instances, GA-BPGnc stops after 61–97 iterations, reaching the objective gap faster than all baselines in iterations and wall-clock time.

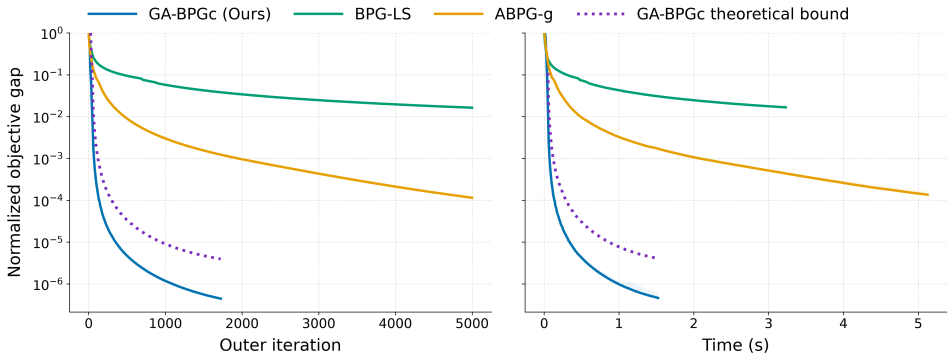

Figure 4: Log plot of the normalized objective gap for D-optimal design on the `abalone` data.

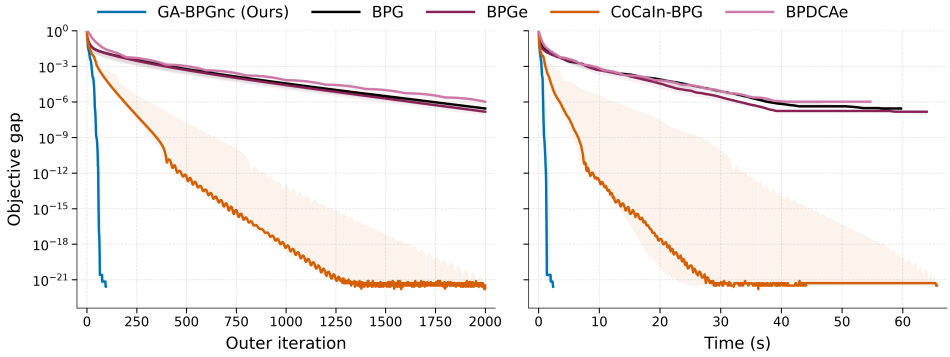

Figure 5: Log plot of the normalized objective gap for nonconvex ℓ_1-regularized phase retrieval.

Nonconvex linear inverse problem. We solve $\min_{x\in\mathbb{R}^n} \Phi(x) := \sum_{i=1}^m \log\big(1 + ((Ax-b)_i)^2\big) + 10^{-3}\|x\|_1$, where $A \in \mathbb{R}^{m\times n}$ has independent $\mathcal{N}(0, 1/m)$ entries ($m = 1000$, $n = 2000$) and $b = Ax^\star + \varepsilon$. The vector $x^\star$ has 10% nonzero entries drawn independently from $\mathcal{N}(0, 50^2)$, and ε is Gaussian noise with standard deviation 10^{-3}. The smooth term has a $2\|A\|_2^2$-Lipschitz gradient, where $\|A\|_2$ denotes the spectral norm. Figure 6 compares GA-BPGnc, BPG, BPGe, and CoCaIn-BPG using Euclidean geometry and $2{,}000$ outer iterations from $x_0 = 0$. We report medians and interquartile ranges of the normalized empirical objective gap over five independently generated $(A, x^\star, \varepsilon)$ instances; Φ_{ref} is each instance's minimum objective across $5{,}000$-iteration runs of all four methods.

F.2 Discussion

In the convex and relatively strongly convex experiments, GA-BPGc and GA-BPGsc outperform nonaccelerated BPG-LS and accelerated ABPG-g despite using smaller median proximal stepsizes.

GA-BPGnc also outperforms CoCaIn-BPG, which uses adaptive stepsizes, in all three nonconvex experiments. Ablation experiments would be needed to measure the separate contributions of stepsize adaptation, extrapolation, and resets.

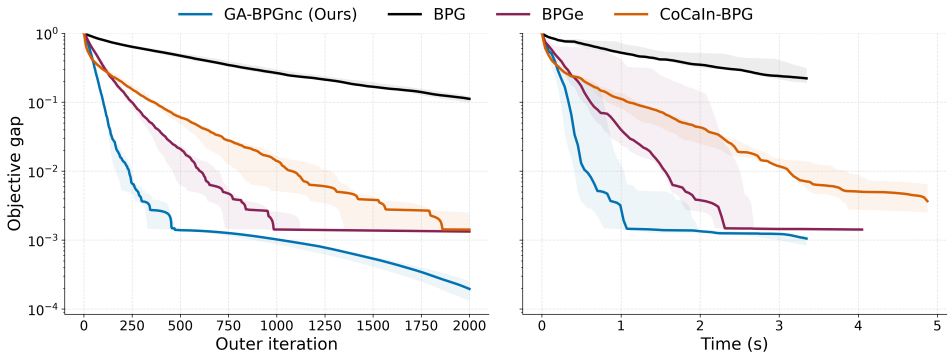

Figure 6: Log plot of the normalized objective gap for the nonconvex linear inverse problem.